THESIS

DOCTORAL PROGRAMME IN INFORMATION AND COMMUNICATION TECHNOLOGIES

Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages

From English to Guarani

Author
Marvin Matías Agüero Torales

Advisor
Antonio Gabriel López Herrera

The School of Technology and Telecommunications Engineering

Granada January, 2022

Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages

From English to Guarani

Author
Marvin Matías Agüero Torales

Advisor
Antonio Gabriel López Herrera

Editor: Universidad de Granada. Tesis Doctorales
Autor: Marvin Matías Agüero Torales
ISBN: 978-84-1117-233-2
URI: http://hdl.handle.net/10481/72863

Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages: From English to Guarani

Marvin Matías Agüero Torales

Keywords: text mining, natural language processing (NLP), machine learning, deep learning, code-switching, low-resource languages

Abstract

This dissertation has focused on the study of machine learning techniques for sentiment analysis and topic modeling in texts from social media. It puts a special emphasis on approaches and methods for handling low-resource languages, i.e., languages lacking large monolingual or parallel corpora and/or manually elaborated linguistic resources sufficient for building Natural Language Processing (NLP) applications; and the implementation of these approaches and methods to multilingual scenarios, such code-switching (i.e., alternating between two or more languages or varieties of language in a phrase or word).

First, we presented a data science workflow to perform machine learning models for social media texts written in low-resource languages, even if these suffer code-switching. The workflow proposed is able to handle different difficulties for the purpose at hand (such as, for example, web text collection, dealing with unbalanced classes, or implementing cross-lingual models).

In the following, we described how to build machine learning models to perform topic modeling with large data coming from social media with short texts written in Spanish, as well as a number of sentiment analysis related tasks for Guarani (a South American indigenous language) and Jopara (i.e., Guarani-Spanish mixture), namely polarity classification, emotion recognition, humor detection, and offensive and toxic language identification. Emphasis was also placed on noisy and short texts coming from social media.

Experiments with the corpora created and the evaluation of the machine learning models built, show the robustness of the approaches and methods proposed in this dissertation, in monolingual, multilingual, and code-switching settings.

The contributions presented in this dissertation may be useful both for the Spanish-speaking community and the Guarani-speaking community. There are many use cases in different areas and disciplines that can benefit from the insights created by the approaches we presented in this thesis. Therefore, there are a number of possible applications for the democratization of low-resource languages, such as the ability to perform less biased monitoring of social networks in multilingual environments or the capacity to extract automatically the knowledge available in non-dominant languages.

Enfoques de aprendizaje automático para el análisis de sentimientos y temas en opiniones multilingües y en idiomas con escasez de recursos: Del inglés al guaraní

Marvin Matías Agüero Torales

Palabras claves: minería de texto, procesamiento de lenguaje natural (PLN), aprendizaje automático, aprendizaje profundo, code-switching, idiomas con escasez de recursos

Resumen

Esta tesis se ha centrado en el estudio de técnicas de aprendizaje automático para el análisis de sentimientos y el modelado de temas en textos procedentes de medios sociales. Se ha puesto un énfasis especial en los enfoques y métodos para el manejo de idiomas conocidos como *low-resource*, es decir, lenguas que carecen de grandes corpus monolingües o paralelos y/o de recursos lingüísticos elaborados manualmente suficientes para construir aplicaciones de Procesamiento del Lenguaje Natural (PLN); y en la aplicación de estos enfoques y métodos en escenarios multilingües, como el *code-switching* (es decir, alternar dos o más lenguas o variedades lingüísticas en una frase o palabra).

Por un lado, introdujimos un flujo de trabajo de ciencia de datos para llevar a cabo modelos de aprendizaje automático para textos provenientes de medios sociales y escritos en idiomas *low-resource*, incluso si estos presentan *code-switching*. Este flujo de trabajo es capaz de lidiar con diferentes dificultades presentes en este ámbito (como, por ejemplo, la colección de texto en la web, el tratamiento de clases con ejemplos desequilibrados o la implementación de modelos multilingües).

Por otra parte, describimos cómo construir modelos de aprendizaje automático para realizar modelado de temas con datos masivos provenientes de medios sociales, con textos cortos en español, así como una serie de tareas para el análisis de sentimientos en guaraní (una lengua indígena sudamericana) y jopara (es decir, la mezcla del guaraní con el español), a saber, la clasificación de polaridad, el reconocimiento de emociones, la detección de humor y la identificación de lenguaje ofensivo y tóxico, también con énfasis en los textos cortos y gramaticalmente pobres provenientes de las redes sociales.

Los experimentos con los corpus creados y la evaluación de los modelos de aprendizaje automático construidos, muestran la robustez de los enfoques y métodos propuestos en esta tesis, tanto en entornos monolingües y multilingües, como de *code-switching*.

Las aportaciones presentadas en esta tesis pueden ser útiles tanto para la comunidad hispanohablante como para la comunidad guaraní-hablante. Hay muchos casos de uso en diferentes áreas y disciplinas que pueden beneficiarse de las ideas creadas por los enfoques que proponemos aquí. Por lo tanto, existen una serie de posibles aplicaciones para la democratización de las lenguas de bajo recursos, como la capacidad de realizar un seguimiento menos sesgado de las redes sociales en entornos multilingües o la capacidad de extraer automáticamente el conocimiento disponible en los idiomas no dominantes.

Antonio Gabriel López Herrera, Associate Professor of the Department of Computer Science and Artificial Intelligence of the University of Granada.

Reports:

That the present work, titled ***Machine Learning approaches for Topic and Sentiment Analysis in multilingual opinions and low-resource languages: From English to Guarani***, has been carried out under his supervision by **Marvin Matías Agüero Torales**, and he authorizes the defense of said work before the corresponding tribunal.

And for the record, he issues and signs this report in Granada on January, 2022.

The advisor:

Antonio Gabriel López Herrera

The doctoral candidate **Marvin Matías Agüero Torales** and the thesis advisor **Antonio Gabriel López Herrera**.

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis advisor and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Place and date: Granada on January, 2022.

Thesis advisor:	Doctoral candidate:
Signature	Signature

I would like to dedicate this thesis...

To my daughter. Dear daughter, you are the greatest gift of my life. May that joy that characterizes you never fade. And may you never tire of fighting for what you love most.

To my parents. Despite the distance, for the infinite support they have given me during all these years of intense periods.

To my wife. Who always is there when I need her. Without her understanding and motivation all this time, I would never have been able to do it.

Acknowledgements

Needless to say, this thesis would not have been possible without the support of many people. First of all, I would like to thank my advisor, Antonio Gabriel López Herrera. He provided me with a great balance between guidance and freedom, helping me to grow professionally. Thank you for initiating me in this exciting field of research by sharing your knowledge, I will always be grateful to you. Thank you for your valuable comments on the thesis. Antonio, thank you for all the talks throughout these years, in which I have been able to learn about research and many other topics in life.

I am also grateful for the helpful contributions of all the co-authors of the publications of this thesis. Especially David Vilares Calvo (Universidade da Coruña) and José Ignacio Abreu Salas (Universitat d'Alacant). David is one of the most brilliant NLP researchers I know and I am grateful for having had the opportunity to collaborate with him. Thank you also for reading all the drafts of the papers and for your helpful comments on them. Thank you for your perspective and for helping me pursue and define more impactful work. José showed up at the right time, helped me with several things, and was able to make a great collaboration possible thanks to his great analytical and synthesis skills, as well as always giving me his helpful comments, support and encouragement. I thank both of them for always being there, even though work time does not always allow it. I would also like to thank Antonio Miranda-Escalada, Jorge Saldívar and Juan Abasolo for the always fruitful discussions.

But the greatest thanks for this thesis go to my family. To my wife and my daughter, for letting me steal that time with them at the rate of hours with the 'computer', and for all the sacrifice that this may represent and all that we have gone through to get here. I love you. To my parents and my brothers, especially, both my mom and dad, for always supporting me, for their invaluable advice, for teaching me to be constant, about the importance of studying, about hard work and its fruits. Thank you mom and dad.

I would be ungrateful if I did not give a special 'thank you' to the friends who always show up in the middle of the night to help in whatever is needed. They listen to you and with a good conversation they make you forget your problems. Thanks also to the new friends I have made during this thesis.

During these years, I have had to combine my doctoral studies with full-time jobs and, in a way, this research has been possible thanks to this financial support. So I would like to express my sincere thanks to the companies, universities and research institutes that hosted me during this time.[1] I would especially like to thank Susana Ladra González (UDC) for trusting me, as well as Diana Roig-Sanz (UOC).

[1] **Paraguay**: Banco Atlas, Banco BASA and UCOM - Universidad Comunera. **Spain**: Universidade da Coruña (UDC), Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS), Universidad de Cádiz (UCA) and UOC (Universitat Oberta de Catalunya).

Contents

Chapter 1

Introduction and Preliminaries

The chapter briefly introduces this thesis and aims to provide an overview of the background relevant to our research. It sets out the hypothesis and its justification, describing the specific problems we address. Furthermore, it describes the general objectives set out to address these problems and the methodology used throughout the dissertation. Finally, the structure of this dissertation is presented.

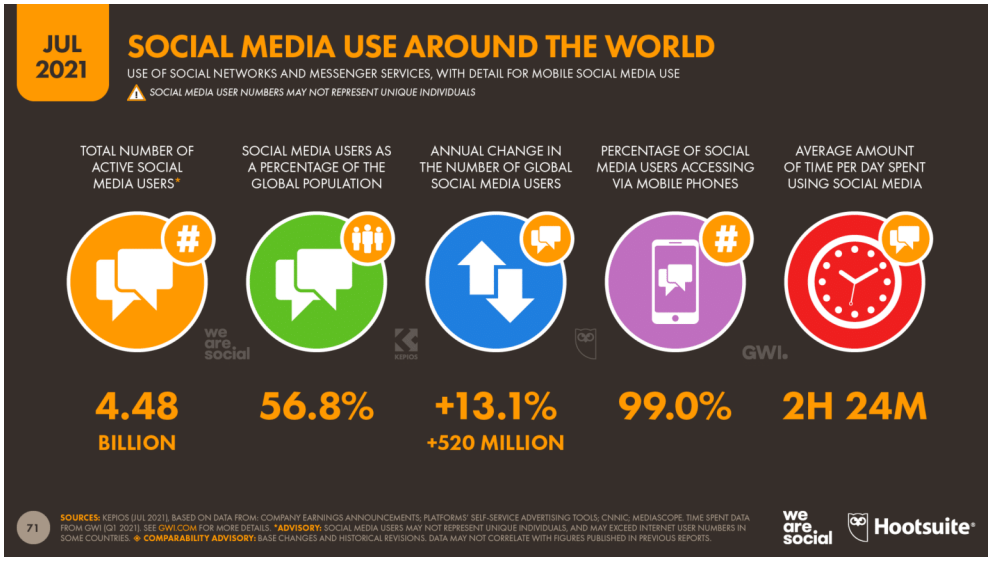
1.1 Introduction

In recent years, the Internet has become the primary source of information, especially social media, where people can be seen sharing their opinions, stances, feelings and experiences. These are writing blog posts, contributing to discussion forums, posting comments on microblogging (e.g., Twitter),[1] interacting with social networks (e.g., Facebook),[2] reviewing a product, service, place, movie, etc. (e.g., Amazon,[3] TripAdvisor,[4] Booking,[5] IMDb[6] or any other social media). That is, these so-called 'social media' have adapted to users, moving from their role as mere consumers of contents, to also producers of information. The opinions and sentiments expressed in a text are especially useful in determining the writer's attitude towards a topic, product, service, etc. For example, this can benefit companies to better position their brand, as well as their products and services. In turn, this type of information also benefits people in their mental decision-making process when buying a product or service, on the basis of what other people have previously written about them.

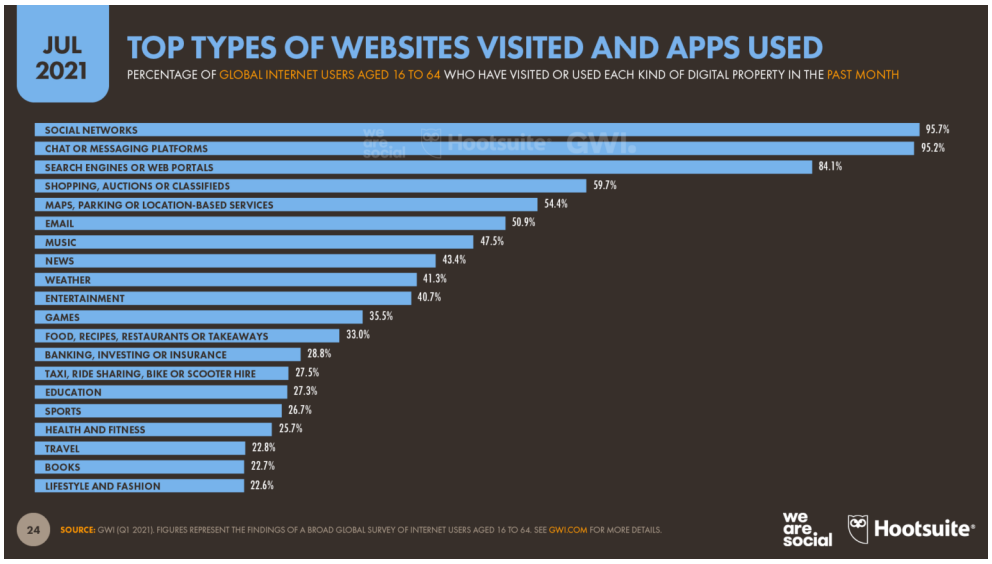
As a result, the World Wide Web offers a huge amount of publicly available textual content, as well as other types of content such as images, videos or audios. Social media continues to drive connected activity worldwide (according data from July 2021 [1]): there are currently 4.8 billion Internet users worldwide (equivalent to almost 61% of the world's population), of which 4.48 billion are social media users (almost 57% of the total population), see Figures 1.1a and 1.1b, respectively. Thus we can see that such a large number of active users will in turn be able to produce a vast amount of content on the web [1] (see Figure 1.2).

The analysis of web content generated by all these users has attracted researchers from different fields, more particularly, Natural Language Processing (NLP). This focuses on the application of computational techniques for the analysis and synthesis of natural language,[7] and more specifically in the case of the web, in its text-rich content.

[1] `https://twitter.com/`
[2] `https://www.facebook.com/`
[3] `https://www.amazon.com/`
[4] `https://www.tripadvisor.com/`
[5] `https://www.booking.com/`
[6] `https://www.imdb.com/`
[7] From `https://www.lexico.com/definition/natural_language_processing`.



(a) Social media use around the world.



(b) Top types of Websites visited and Apps used.

Figure 1.1: Social media statistics [1].

Besides, these online opinions, comments, reviews, i.e. textual contents can be expressed in different languages, even mixing them in a sentence or a word. This practice is called code-switching or code-mixing, as we can see for example in Jopara, a mixture of Guarani (a South American indigenous language) and Spanish.[8] Therefore, it is necessary to analyze all these textual contents in all possible languages and dialects. In addition, there are languages that are worse off than their peers in terms of linguistic resources (e.g., pre-trained models, lexicons or annotated corpora), the so-called low-resource languages. On the other hand, these rich-resource languages, such as English or Chinese, are considered high-resource. For example, Spanish is widely spoken in the world [2], however, especially when compared to those mentioned above, it lacks similar resources.

Analyzing each language independently requires enormous human effort and is a hard, long and expensive task, so the so-called rich-resource languages have the advantage against the low-resource languages. There is a clear mismatch between the distribution of resources

[8]From `https://www.lexico.com/definition/code-switching`.

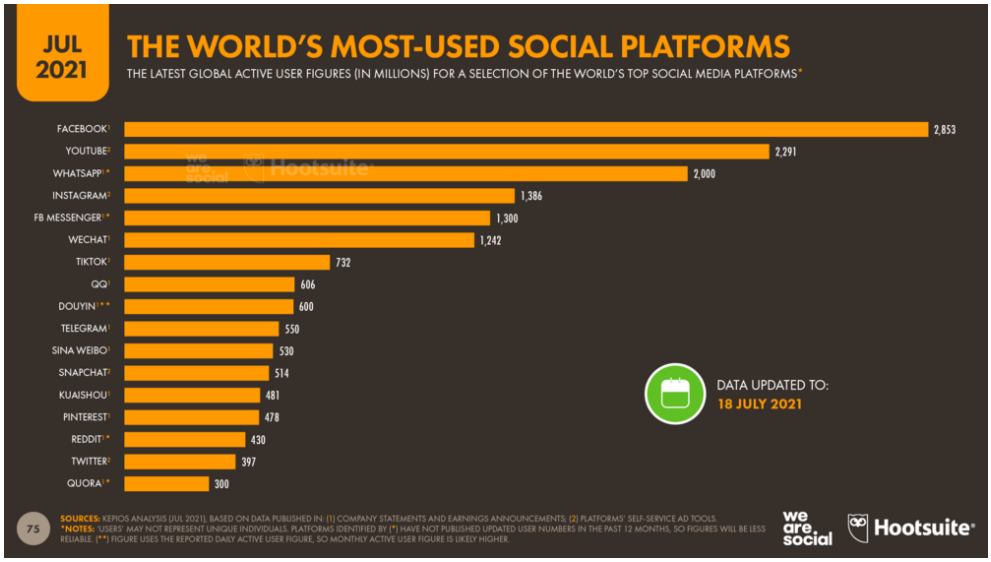


Figure 1.2: World's most used social media platforms [1].

among the different languages. It is necessary to democratize these language technologies building resources, models or systems capable of analyzing all languages, without forgetting the worst positioned ones.

Thus, this thesis proposes machine-learning approaches to deal with opinion analysis mainly in low-resource languages for monolingual, multilingual and code-switching environments for both sentiment analysis and topic modeling. More particularly, Spanish, Guarani and Jopara. While sentiment analysis aims to determine the sentiment in a text, topic modeling attempts to discover the topics that can be frequently found in a collection of texts. Our goal is to study the different existing machine learning approaches for these tasks, as well as to develop new linguistic resources for text analysis in resource-poor languages (such as corpora, annotation guides, software tools, etc.), especially those written in social media.

After this brief introduction, the remainder of this chapter is organized as follows: the section 1.2 aims to provide an overview of the background relevant to our research. The section 1.3 and the section 1.4 then set out the hypothesis and justification for this manuscript, describing the particular problems we address. The general objectives followed in this thesis to address these problems are described in the section 1.5. In the section 1.6, the methodology used throughout the dissertation is described. Finally, in the section 1.7, the structure of this thesis dissertation is presented.

Introducción

En los últimos años, Internet se ha convertido en la principal fuente de información, especialmente las redes sociales, donde se puede ver a la gente compartiendo sus opiniones, posturas, sentimientos y experiencias. Se trata de escribir entradas en blogs, contribuir a foros de debate, publicar comentarios en microblogging (por ejemplo, Twitter),[9] interactuando con las redes sociales (por ejemplo, Facebook),[10] reseñando un producto, servicio, lugar, película, etc. (por ejemplo, Amazon,[11] TripAdvisor,[12] Booking,[13] IMDb[14] o cualquier otro medio social). Es decir, estos llamados 'medios sociales' se han adaptado a los usuarios, pasando de su papel de meros consumidores de contenidos, a ser también productores de información. Las opiniones y sentimientos expresados en un texto son especialmente útiles para determinar la actitud del escritor hacia un tema, producto, servicio, etc. Por ejemplo, esto puede beneficiar a las empresas para posicionar mejor su marca, así como sus productos y servicios. A su vez, este tipo de información también beneficia a las personas en su proceso mental de toma de decisiones a la hora de comprar un producto o servicio, basándose en lo que otras personas han escrito sobre ellos previamente.

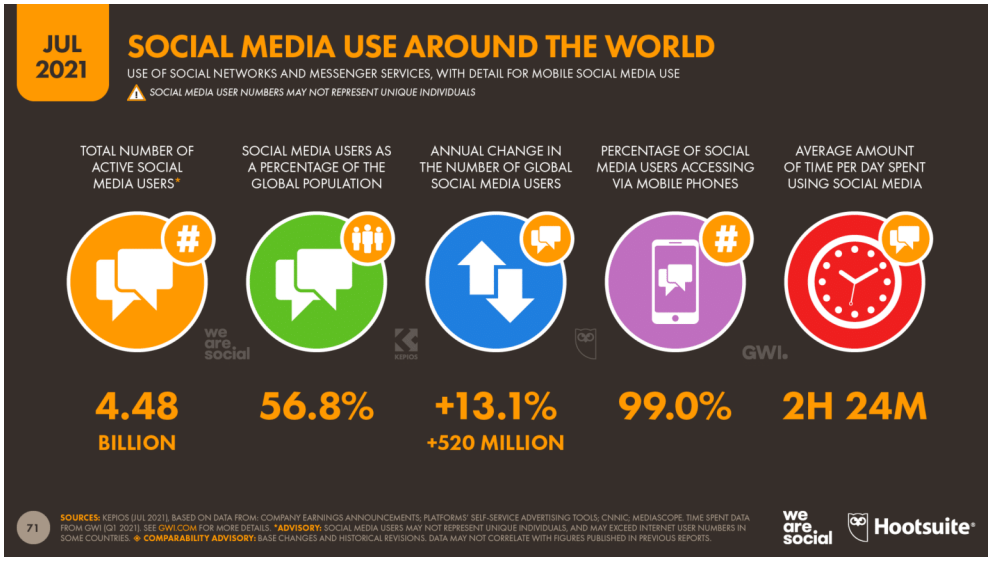
Como resultado, la *World Wide Web* ofrece una enorme cantidad de contenido textual disponible públicamente, así como otros tipos de contenido como imágenes, vídeos o audios. Los medios sociales siguen impulsando la actividad conectada en todo el mundo (según datos de julio de 2021 [1]): actualmente hay 4,800 millones de usuarios de Internet en todo el mundo (lo que equivale a casi el 61% de la población mundial), de los cuales 4,480 millones son usuarios de medios sociales (casi el 57% de la población total), véanse las Figuras 1.3a y 1.3b, respectivamente. Por lo tanto, podemos ver que un número tan grande de usuarios activos podrá, a su vez, producir una gran número de contenidos en la web [1] (véase la Figura 1.4).

El análisis de los contenidos web generados por todos estos usuarios ha atraído a investigadores de diferentes campos, más particularmente, del Procesamiento del Lenguaje Natural (PLN). Este se centra en la aplicación de técnicas computacionales para el análisis y la síntesis del lenguaje natural,[15] y más concretamente en el caso de la web, en su contenido rico en texto.

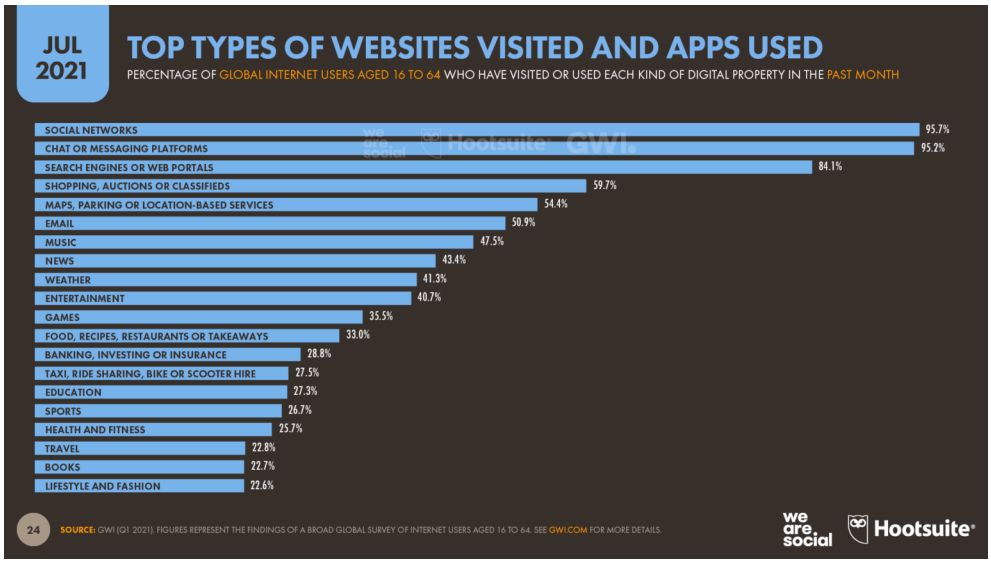
Además, estas opiniones, comentarios y críticas en línea, es decir, los contenidos textuales, pueden expresarse en diferentes idiomas, incluso mezclándolos en una misma frase o palabra. Esta práctica se denomina *code-switching* o *code-mixing*, como podemos ver por ejemplo en el *Jopará*, una mezcla de guaraní (una lengua índigena sudamericana) y español.[16] Por lo tanto, es necesario analizar todos estos contenidos textuales en todas las lenguas y dialectos posibles. Así, hay lenguas que están en peor situación que sus pares en cuanto a recursos lingüísticos (por ejemplo, modelos preentrenados, léxicos o corpus anotados), las llamadas lenguas de bajos recursos, también conocidas como *low-resource*. En cambio, las lenguas ricas en recursos, como el inglés o el chino, se consideran de altos recursos. Por ejemplo, el español es un idioma muy hablado en todo el mundo [2], sin embargo, especialmente si se compara con las mencionados anteriormente, carece de recursos equiparables.

Analizar cada lengua de forma independiente requiere un enorme esfuerzo humano y es una tarea ardua, larga y costosa, por lo que las denominadas lenguas ricas en recursos tienen ventaja frente a las de bajos recursos. Existe un claro desajuste entre la distribución de

[9] `https://twitter.com/`
[10] `https://www.facebook.com/`
[11] `https://www.amazon.com/`
[12] `https://www.tripadvisor.com/`
[13] `https://www.booking.com/`
[14] `https://www.imdb.com/`
[15] De `https://www.lexico.com/definition/natural_language_processing`.
[16] De `https://www.lexico.com/definition/code-switching`.



(a) Uso global de los medios sociales.



(b) Principales tipos de sitios web visitados y aplicaciones utilizadas.

Figure 1.3: Estadísticas de los medios sociales [1].

recursos entre las distintas lenguas. Es necesario democratizar estas tecnologías lingüísticas mediante la construcción de recursos, modelos o sistemas capaces de analizar todas las lenguas, pero sin olvidar las peor posicionadas.

Así pues, esta tesis propone enfoques de aprendizaje automático para abordar el análisis de opiniones principalmente en lenguas con ezcasez de recursos para entornos monolingües, multilingües y de *code-switching*, tanto para el análisis de sentimientos como para el modelado de temas. En concreto, el castellano, el guaraní y el jopará. Mientras que el análisis de sentimientos tiene como objetivo determinar el sentimiento en un texto, el modelado de temas intenta descubrir los temas que se pueden encontrar frecuentemente en una colección de textos. Nuestro objetivo es estudiar los diferentes enfoques de aprendizaje automático existentes para estas tareas, así como desarrollar nuevos recursos lingüísticos para el análisis de textos en idiomas con pocos recursos (como corpora, guías de anotación, herramientas informáticas, etc.), especialmente los escritos en medios sociales.

Luego de esta breve introducción, el resto del capítulo se estructura de la siguiente man-

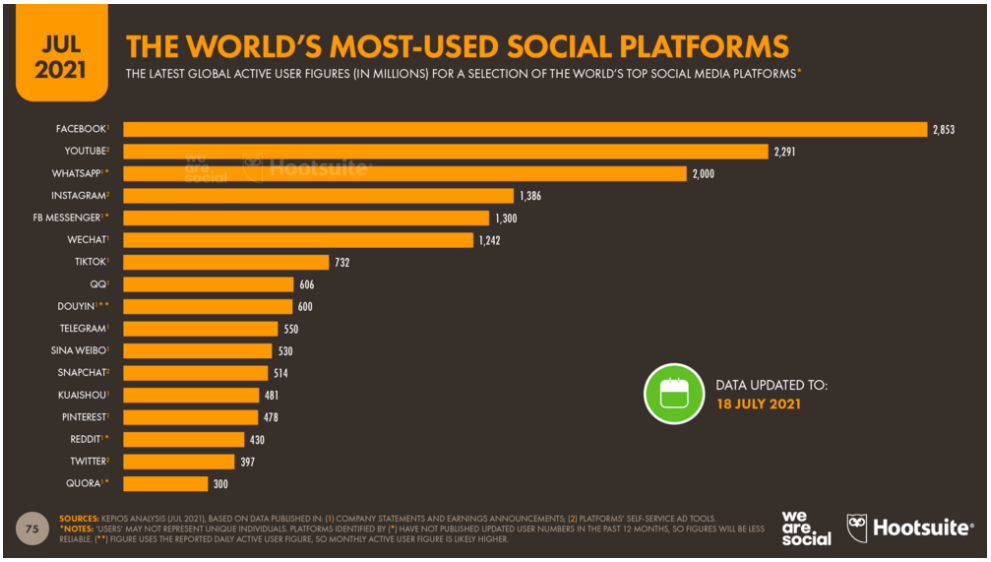


Figure 1.4: Plataformas de medios sociales más utilizadas del mundo [1].

era: la sección 1.2 tiene como objetivo proporcionar una visión general de los antecedentes relevantes para nuestra investigación. A continuación, las secciones 1.3 y 1.4 exponen la hipótesis y la justificación de este manuscrito, describiendo los problemas particulares que planteamos. Los objetivos generales que se siguen en esta tesis para abordar estos problemas se describen en la sección 1.5. En la sección 1.6, se describe la metodología utilizada a lo largo de la tesis. Por último, la sección 1.7 presenta la estructura de la misma.

1.2 Preliminaries

The automatic processing of opinions is possible thanks to the computational techniques available, which in turn are due to artificial intelligence together with natural language processing and the advancement of the field of machine learning. These opinions are processed computationally, identifying and categorizing or grouping the opinions expressed in a text, document, or collection of these. Without machine learning it would not be possible to perform these NLP tasks with near-human accuracy, at least without investing considerable time, although performing these tasks manually is impossible when dealing with large-data. This section places the contributions of this thesis in a broader perspective, as the problem of multilingualism, i.e., that not only can opinions written in one language exist, but they can also be written in several languages, even at sentence or document level, and these languages, in turn, can belong to those considered as low-resource, which leads to the need for the development of resources, approaches, and solutions for this particular type of problems, where machine learning and natural language processing proved to be really useful.

In this section, we give an overview of the concepts relevant to the contributions of this thesis. We present an overview of Machine Learning in subsection 1.2.1. In subsection 1.2.2, we introduce the concept of low-resource languages. The subsection 1.2.3 discuss terminology related to sentiment analysis and comment on some involved problems, and subsection 1.2.4 gives a brief overview of topic modeling. Finally, subsection 1.2.5 provides a general view of the NLP workflow and of the most relevant tasks of Natural Language Processing, as well as its history and evolution.

1.2.1 Machine Learning

Machine Learning (ML), a sub-field of study of Artificial Intelligence (AI), refers to the use and development of computer systems capable of 'learning', adapting without following explicit instructions, by using algorithms and mathematical and statistical models to analyze and draw conclusions from patterns in data.[17] It relates to the design, analysis, optimization, development and application of such methods. Machine learning usually consists of the basis of two large phases:

1. Model estimation: The first phase of estimating a model from data, or 'training' phase, is usually performed before the practical use of the model, consists of solving a practical task, such as translating a text from a language to another, recognizing the presence of an object in a photograph or video, and so on.

2. Model implementation: The second phase corresponds to the production phase once the model has been determined, new data can be presented to obtain the result corresponding to the desired task (i.e., the 'prediction' phase). It should also be remembered that some systems can continue learning in this phase, provided they have a means of obtaining feedback on the quality of the results produced.

Machine learning can be categorized in different ways, according to the information available during the training phase of the model and in a general way [3]:

- Supervised learning: Occurs when the data are previously labeled or annotated (i.e., the answer to the task for that data is known). It is called 'classification' if the labels are discrete, or 'regression' if they are continuous.

- Unsupervised learning: In this case, there are no labels, so the underlying structure of the data (which may be a probability density or a clustering, for example) is pursued.

[17] From `https://www.lexico.com/definition/machine_learning`.

- Semi-supervised learning: Here the two previous types of algorithms are combined to generate a suitable model. It is characterized by using a combination of labeled and unlabeled data in the training phase.

- Reinforcement learning: It is considered as reinforcement learning if the model is learned incrementally based on a reward received by the program for each action performed.

To recap, machine learning models are trained with some learning data, labeled or unlabeled, using techniques and algorithms to predict them, with or without additional data. The most notable approaches to create models for machine learning systems are as follows:

- Artificial Neural Network: An artificial neural network, usually called a 'neural network' (NN) or simply a neural model, is an algorithm inspired by biological neural networks, i.e. in its structure and functional aspects.[18] In neural networks, computations are structured in terms of an interconnected group of artificial neurons, which process information using a connection computation approach with numerical weights.

  - Similarly, *Deep Learning (DL)* consists of multiple input, output and hidden layers in an artificial neural network.[19] Therefore, these networks can create multiple levels of abstraction that represent data. This allows a machine to train itself to perform a task. For example, they are often used to model complex relationships between inputs and outputs, to find patterns in data, among others. So, may refer to [4] for more details.

- Bayesian networks: In a Bayesian network, a probabilistic graphical model, each node corresponds to a random variable and each edge represents the conditional probability for the corresponding random variables, which is used for representing knowledge about an uncertain domain [5]. The Naive Bayes model is the simplest and commonly used Bayesian network model.

- Clustering: Clustering is a common technique and a common starting point in statistical data analysis. Clustering analysis consists of assigning a set of 'clusters' such that items within the same cluster are similar according to some pre-established criteria, while items in different clusters, are not similar [6].

- Decision trees: Decision tree learning uses the decision tree as a prediction model [7]. As a decision tree is a tree diagram used to represent the different stages of a decision-making process, decision tree learning relates those observations about an element to conclusions about the objective value of that element.

- Genetic algorithms: A genetic algorithm is an evolutionary algorithm, which mimics the process of natural selection, thereby handling the possible solutions to a given problem numerically.[20]

- Support Vector Machines: Support Vector Machines (SVMs) are a set of supervised learning methods [8], which given a set of training examples, each marked as belonging to one or more categories, build a model that predicts whether a new example fits one category or another. SVMs can be used for both classification and regression.

[18] From `https://www.lexico.com/definition/neural_network`.
[19] From `https://www.lexico.com/definition/deep_learning`.
[20] From `https://www.lexico.com/definition/genetic_algorithm`.

1.2.2 Low-resource languages

Natural language processing for resource-poor languages, also known as low-resource, is one of the great problems that are still unsolved in today's digital society, given the difficulties of creating accurate models due to the scarcity of annotated data [9]. In this thesis, computational models will be used for the development of natural language processing technologies capable of automatically understanding the sentiments expressed by users or discovering topics in social media in languages known as low-resource languages. That is, those with few speakers or worse positioned (e.g., economic inequalities, absence of resources, languages which are mostly oral).

The reasons for the importance of this field are various [9, 10]: from having the ability to automatically extract knowledge only available in non-dominant languages or being able to perform less biased social network monitoring in multilingual environments (e.g. in Paraguay, where Paraguayan Spanish formally predominates but Guarani[21] [22] and Jopara,[23] prevails in 'word of mouth'). Even to ensure democratic access to linguistic technologies such as the summarization of opinions and the topics on which they are based. More particularly, one of the objectives of the former is to be able to give the user all the information already processed, showing him only what he is interested in, based on the opinion of hundreds of users who have shared their experience. This also includes that society can immediately know a summary of what people think about topics of public interest, such as certain political measures, local problems or events [10].

Table 1.1 shows four categories of low-resource languages and some examples of them highlighting their characteristics and language families, according to [11], where (1) high-resource languages and (2) low-resource languages, are those covered by the SOTA (*State Of The Art*) multilingual models (i.e., by Multilingual BERT 'mBERT'[24] [12] and Cross-lingual Language Model - XLM-R [13], both being trained on large-data of about 100 or more languages). On the other hand, (3) low-resource languages and (4) truly low-resource languages are those not covered by the above multilingual models. Furthermore, for this categorization of languages in [11], they take into account the variation in data availability related to the size of their Wikipedias.

In this thesis we work with two languages with few resources compared with English: first (i) Spanish (see chapter 3 and chapter 4) and then with (ii) Guarani (see chapter 3, chapter 5 and chapter 6). In addition, we have worked with a code-switching language between the last two (i.e., Guarani mixed with Spanish, called 'Jopara'). On the one hand, Spanish does not belong to the extremely low resource category according to [14], hence we have categorized it as (2) in Table 1.1. Guarani, on the other hand, is a 'truly low-resource language', as can be seen in Table 1.1, with very few linguistic resources indeed.

1.2.3 Sentiment Analysis

One of the most popular tasks that NLP systems solve is commonly known as Sentiment Analysis (SA) [15]. The main idea is that given a text, a built model must be able to determine its sentiment or emotion. This sentiment can be of positive to negative polarity, or an emotional state such as happy, angry or sad [16, 17]. When performed on a rich-resource language, it is solved in a relatively simple way with models that could be considered as traditional (e.g., Naive Bayes or Support Vector Machine - SVM). But to do it in a

[21] Guarani is a South American indigenous language spoken mainly in Paraguay, northeastern Argentina, southern Bolivia and central-western Brazil, bordering Paraguay.

[22] Guarani is also an official language of Mercosur, a common market organization that involves a number of Latin American countries. From `https://normas.mercosur.int/simfiles/normativas/10443_DEC_035-2006_ES_IncorporaIdiomaGuarani.pdf`.

[23] Jopara ('joparâ' in Guarani), is a mixture of Guarani language and Spanish.

[24] `https://github.com/google-research/bert/blob/master/multilingual.md`

	Language	ISO code	Language family	# of Wiki articles*	Covered by SOTA?
(1)	English	en	Indo-European	6.0M	✓
	Japanese	ja	Japonic	1.2M	✓
	Chinese	zh	Sino-Tibetan	1.1M	✓
	Arabic	ar	Afro-Asiatic	1.0M	✓
(2)	Javanese	jv	Austronesian	57K	✓
	Swahili	sw	Niger-Congo	56K	✓
	Icelandic	is	Indo-European	49K	✓
	Burmese	my	Sino-Tibetan	45K	✓
(3)	Quechua	qu	Quechua	22K	
	Min Dong	cdo	Sino-Tibetan	15K	
	Ilokano	ilo	Austronesian	14K	
	Mingrelian	xmf	Kartvelian	13K	
(4)	Meadow Mari	mhr	Uralic	10K	
	Maori	mi	Austronesian	7K	
	Turkmen	tk	Turkic	6K	
	Guarani	gn	Tupian	4K	

*M: Million, K: thousand.

Table 1.1: Low-resource languages categories and examples of them in each category [11, p. 7658, Table 1].

multilingual way, i.e. determining the sentiments of texts written in multiple languages, either by taking texts composed of any language, or code-switching (mixing languages in the same text), it is necessary to use more advanced machine learning models, such as those based on neural networks, as its tend to obtain similar results to humans [12, 18]. In any case, in order to solve it, it is posed as a classification problem at the sequence level, since it will be necessary to determine the sentiment from a sequence of characters.

1.2.3.1 Related concepts

The growth of the Internet and the Web gave rise to social media, where a huge number of users post their opinions, reviews and comments on social networks, blogs, forums, specialized websites, etc. With this new data, a lot of systems capable of solving this task started to be developed. We can clearly distinguish three levels of analysis that can be carried out [15]: (i) document level, it refers to the opinion expressed by a document as a whole (assuming that only one entity is referred to in the document); (ii) sentence level, it aims to identify the sentences that make up a text and try to infer the orientation of the opinion they contain; and (iii) aspect and entity level, in this level of analysis the opinion is about a specific entity, so it aims to delve into a specific aspect of that entity.

Sentiment analysis is mainly based on the classification of texts according to the intention with which the author wrote it, i.e., whether his attitude towards a certain topic, entity, etc. is either positive, negative or neutral. Actually, this problem is much more complex than that, because it involves the detection of affect in language [19]. There are main types of analysis [20], namely:

- Polarity or classification of sentiment: Helps to determine what opinion is held about the idea of the text, whether positive, negative or neutral. There are cases in which the opinion is given from different perspectives, so the system must determine how important each one of them is. This is usually straightforward, because there are words that directly indicate what the intention is, but this is not always the case (e.g., the presence of code-switching, slangs or negations in the text).

- Subjectivity classification: This is based on determining whether a text is subjective

or not. Sometimes it can be more complicated than polarity classification, since texts may not be exclusively of one type or the other (e.g., news with negative sentiments).

- Opinion summarization and opinion retrieval: Opinion summarization focuses on extracting the main characteristics (and the sentiments related to them) of an entity in one or several documents (i.e., single-document or multi-document summarization). Opinion retrieval, on the other hand, tries to retrieve a ranking of documents that express an opinion on a given issue about a specific query.

- Sarcasm and irony: The detection of sarcasm and irony tries to detect utterances with ironic and/or sarcastic content, which is very complicated and difficult to solve in NLP, since the figurative and creative nature of sarcasm and irony is a great challenge for the so-called affective computer systems [21].

- Other types, to name just a few, include:

  - Emotion recognition: Emotion analysis, recognition or detection are systems that attempt to identify richer differences in sentiment expression in utterances than the mere polarity of the sentence itself. Emotion class labels are often based on (i) Ekman's six basic emotions[25] [22]; (ii) Plutchik's wheel with eight primary emotions [23]; (iii) Parrott's emotions groups [24]; (iv) Scherer's categories of affect [25]; or (v) a set of application-specific emotion-categories may be defined (e.g., [17, 26]).
  - Stance detection: Stance detection focuses on inferring the text author's point of view by linking stance to three factors: linguistic acts, social interactions, and individual identity [27]. It is a classification task in which the stance of the text writer is desired in the form of a category label from a given set, namely, *Favor*, *Against*, *Neither*, and occasionally, *Neutral* [28]. Note that the target may or may not be explicitly mentioned in the text.
  - Humor and funny detection: Humor detection and analysis aim to identify whether a text is funny or not, as well as to estimate the humor of the text on a previously given humor scale, usually numerical [29, 30]. Additionally, recently two new sub-tasks was introduced in [31], which is related to the content of the text: (i) humor logic mechanism and (ii) target classification, where the first one focused on how the joke works and the second one identified the target entity of the joke (if any).
  - Offensive language identification: The identification of offensive, abusive or foul language categories in texts is ubiquitous especially on social media platforms and offensive messages can range from (i) hate speech, which usually consists of insults or swearing directed at a group or individual, and (ii) cyberbullying, which usually targets individuals [32, 33]. Its goal is to classify a text into offensive or non-offensive language, automatic categorization of the types of offense, and, as we saw in humor detection, also the identification of the target of the offense.

It should be noted at this point that the task of sentiment analysis is mostly carried out with data from social media, such as Facebook, Twitter, TripAdvisor, to name a few, and the interest of the scientific community has been growing for some time now [34]. In this thesis we have chosen to work with Twitter (see chapter 3). Note that Twitter is almost always chosen over other social media platforms for the following reasons [35]: (a) Predominance of text-based content: as opposed to other platforms that favor richer and more dispersed content (videos, images, audios, etc.); (b) Diversity of tools for information

[25] anger, disgust, fear, happiness, sadness, and surprise.

extraction: there are almost no restrictions to obtain a meaningful sample of all interactions by providing a set of query parameters (either using the public Twitter search API or by web scraping); (c) High engagement general-purpose platform; and, (d) Existence of geolocated interactions.

1.2.3.2 Multilingual sentiment analysis

The sentiment resources (i.e., lexicons, annotated corpora, pre-trained models, NLP tools and libraries) in non-English languages are difficult to find [36], especially in languages considered as low-resource, due to sentiment analysis has been primarily focused on the English language [37]. Therefore, automatic sentiment analysis systems in other languages (different from the rich-resource ones) vary from language to language and tend to be less accurate. Some work on multilingual sentiment analysis has mainly focused on developing methods, approaches, systems, etc. for transferring sentiment resources from one rich-resource language to another considered as low-resource, known as cross-lingual sentiment analysis [38].

Other approaches have focused on developing language-independent sentiment analysis systems, although the accuracy of a system built following this approach tends to be worse than the effectiveness of monolingual systems, however, this can be useful for languages with truly low resources (e.g., Guarani). This led to the development of multilingual sentiment analysis systems, based on systems such as mBERT [12] or XLM-R [13].

On the other hand, code-switching [39] poses a challenge for sentiment analysis systems, as texts contain terms in two or more different languages, which are intertwined with each other (e.g., Jopara, Guarani-Spanish mixture). In these cases, it is common to use systems such as those mentioned above, multilingual and cross-lingual, or a specific one trained (or pre-trained) in the mixed languages of the target text.

Check chapter 2 for a more exhaustive bibliography of current multilingual sentiment analysis. Figure 1.5 shows an increase of interest of the research community year by year about multilingual sentiment analysis as well as research for low-resource languages sentiment analysis. As of December 1, 2021, we have found 688 papers since 1999 on this issue. In particular, for SA in Guarani, we did not find in the results any previous work; with the exception of [40], who carried out sentiment categorization on a Spanish-dominant Jopara. Note that this is according to the studies published in scientific journals, proceedings, and so on and collected in the Scopus database.[26] It is important to note that, as the year 2021 was not yet closed at the time of the Scopus query (December 1, 2021), we naturally see fewer articles in this year than in the previous one. SA has a wide range of applications and is seen in the Figure 1.6a areas such as computer science, social sciences, business, medicine, etc. For instance, it can be used in medicine to detect suicidal ideation [41]. It can also be used in public opinion detection of political trends [42], in fake news detection [43] or in brand management by electronic word-of-mouth [44].

In Figure 1.6b, authors seem to actually prefer conference over journal submissions. Note that industry interest in conference proceedings is higher because the information is promptly available more than journals.

To perform the small bibliometric analysis shown in Figure 1.5 and 1.6, we chose the Scopus database,[27] as mentioned above, because it offers a more extensive list of modern sources, it also has an independent sourcing system and its interface is easier to use. The query string used for the search is available.[28] We limit ourselves to these publication years

[26] `https://www.scopus.com/`

[27] It is an abstract and citation database by science publisher Elsevier.

[28] (TITLE-ABS-KEY(*multilingual*) OR TITLE-ABS-KEY(*bilingual*) OR TITLE-ABS-KEY(*code-switching*) OR TITLE-ABS-KEY(*code-mix**) OR TITLE-ABS-KEY(*cross-lingual*) OR TITLE-ABS-KEY(*cross-language*) OR TITLE-ABS-KEY(*low-resource**) OR TITLE-ABS-KEY(*poor-resource**) OR

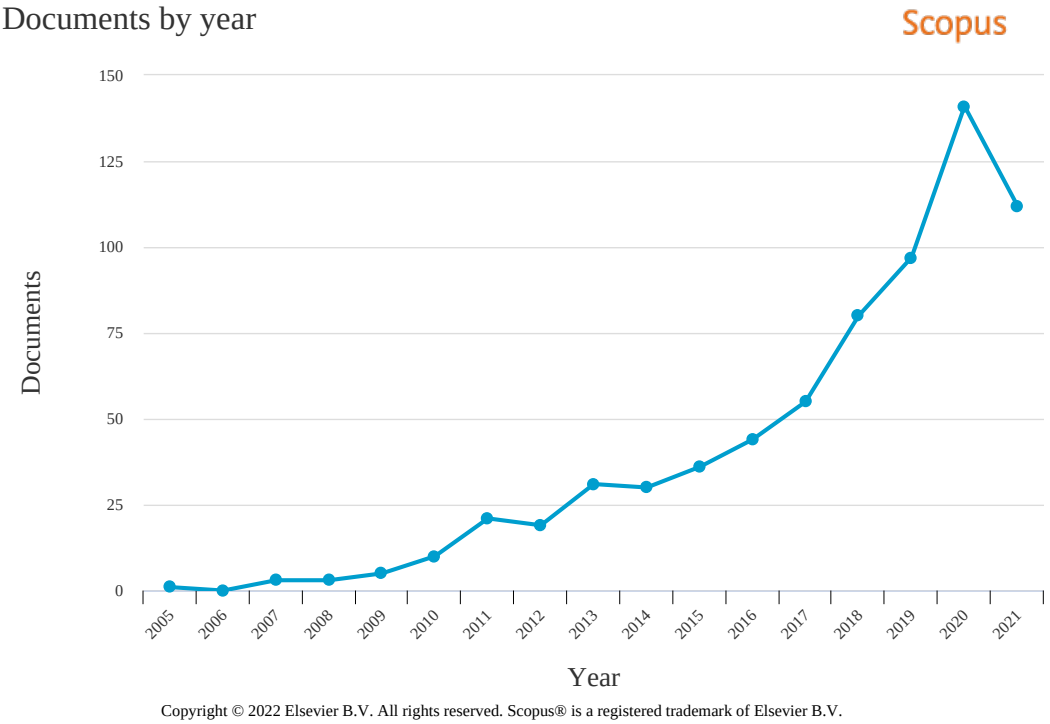


Figure 1.5: Research interest in sentiment and emotion analysis in multilingual and low-resource languages (documents by year). December 1, 2021.

because of the existence, as early as 1999,[29] of work closer to the definition of sentiment analysis, such as that of Wiebe et al. [46], which proposes the development and use of a Gold-Standard dataset for subjectivity classification. Subsequently, the rise of sentiment analysis came with Turney (2002) [47], who based on a dictionary approach, introduced an unsupervised algorithm to compute the semantic orientation of texts [48].

1.2.4 Topic Modeling

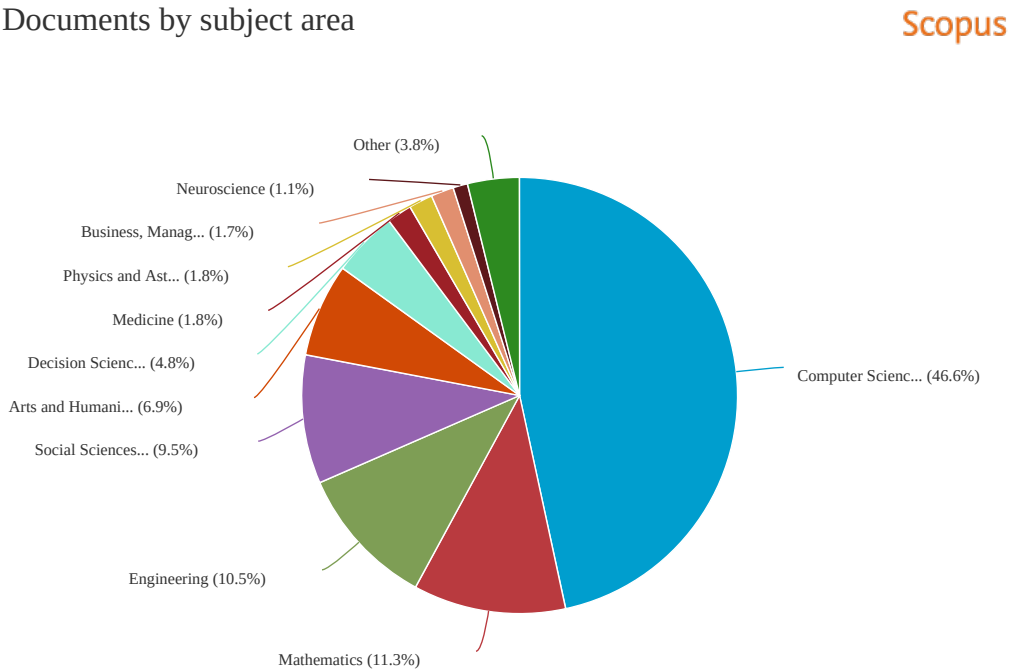
Another NLP's popular task is Topic Modeling or Topic Model, an unsupervised technique for extracting topics by using contextual clues to connect words with similar meanings and to distinguish between uses of words with multiple meanings. Topic model is useful to determine topics that can be frequently found in any text documents. These documents can be written in any language, but English is the most studied, as it has more linguistic resources than its counterparts.

A topic model is a probabilistic model to determine and 'discover' abstract topics in a document, it is useful to know what people are talking about and understand their problems and opinions automatically. Topic modeling is an unsupervised machine learning technique, which requires no training and usually is a useful way to start analyzing large text collections. For instance, these large text collections can be from social media texts, such as reviews of hotels, restaurants, movies, etc., news, emails, state bulletins, tweets, etc. [49].

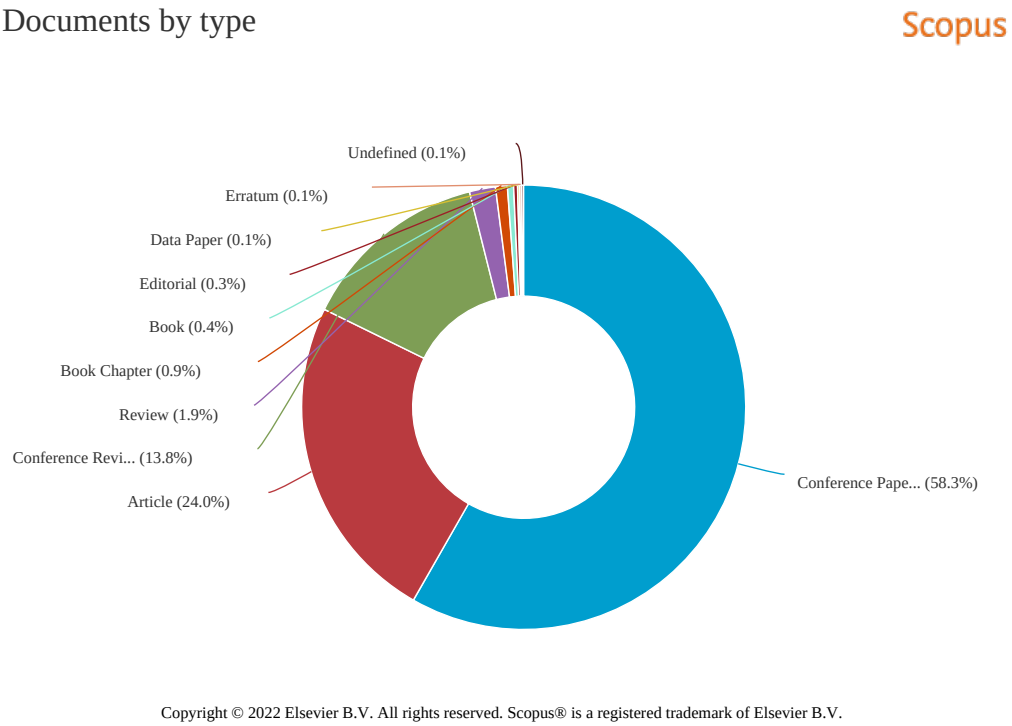
Topic modeling involves counting words and clustering patterns of similar words to infer themes within unstructured data. By detecting patterns such as word frequency and

TITLE-ABS-KEY(*resource-poor**) OR TITLE-ABS-KEY(*"indigenous language"*)) AND (TITLE-ABS-KEY(*"sentiment analysis"*) OR TITLE-ABS-KEY(*"opinion mining"*) OR TITLE-ABS-KEY(*"affect detection"*) OR TITLE-ABS-KEY(*"affective computing"*)) AND (PUBYEAR AFT *1998* AND PUBYEAR BEF *2022*)

[29]Please, refer to Pang and Lee (2008) [45] for a detailed bibliography.



(a) Plot of Scopus search results of papers by subject area.



(b) Plot of Scopus search results of papers by type.

Figure 1.6: Research interest in sentiment and emotion analysis in multilingual and low-resource languages. Documents by (a) subject area and (b) type. December 1, 2021.

distance between words, a topic model clusters comments, opinions, reviews, etc. that are similar and words and expressions that appear more frequently. Thus, with this information, it is possible to deduce what each set of texts is talking about. Therefore, what topic modeling methods do is to try to find out which topics are present in the corpus documents and what is the intensity of that presence. For what a corpus of documents is divided into two (i) a list of the topics addressed by the documents in the corpus and, (ii) several sets of documents in the corpus grouped by the topics they address [50].

The Latent Dirichlet Allocation (LDA), since its development in 2003 [51], is the most common topic model in use, it is a generalization of PLSA (Probabilistic Latent Semantic Analysis) proposed in 1999[30] [53]. Other actual topic models, to a greater extent are extensions of LDA itself [54] or similar alternatives [55]. Although LDA-based topic models have shown good performance, pre-trained neural models are especially useful, as they are supposed to contain more accurate representations of words and sentences [54, 56, 57], which precisely take advantage to create easily interpretable topics, although this is not always the case, like LDA, with some effort through algorithms tuning and refinement, is able to create meaningful and discriminative topics as we will see in chapter 4.

It is difficult to find a large number of papers that perform the task of topic modeling in Spanish or in another language considered to be of limited resources and even more so in large data. This is why it is so important to make contributions in this area. We found only 80 papers since 2003 on this issue. The search query used in Scopus is available.[31] Figure 1.7 shows how intermittent is the interest of the research community on this issue, however, fortunately we see that it has gained prominence. We note in Figure 1.7 the increase of papers on the subject in recent years. It is important to empathize that the few articles in the year 2021 (in relation to the last year) are due to the fact that at the time of the Scopus query (December 1, 2021), this year was not yet closed.

1.2.5 Natural Language Processing

Today, a large amount of information comes from many sources, but the vast majority is through texts written by social media platforms such as social networks, blogs, forums, etc. being necessary to perform automatic language processing. Therefore, a discipline of artificial intelligence called Natural Language Processing (NLP) is growing and has become one of the most important and useful fields within Artificial Intelligence (AI) and Machine Learning (ML), especially Deep Learning (DL). NLP aims to develop systems capable of processing and understanding human language. Since its inception, its potential has been demonstrated in multiple tasks, but only a few years ago, with the growth of Deep Learning, it has positioned itself as a promising discipline.

1.2.5.1 Mini-history of NLP

In 1954 an experiment was conducted at Georgetown University (USA), which together with IBM,[32] automatically translated more than sixty sentences from Russian into English [58]. The authors predicted that within a few years this task should be completely solved, but it was not until much later that systems would be developed that would generate a consistent translation [59]. Indeed, models are still being developed today to perform translations between any language pair, especially those categorized as low-resource. For instance, in

[30] Another earlier model was proposed by Papadimitriou et al. in 2000 [52].

[31] (TITLE-ABS-KEY(*spanish*) OR TITLE-ABS-KEY(*low-resource**) OR TITLE-ABS-KEY(*resource-poor**) OR TITLE-ABS-KEY(*poor-resource**)) AND (TITLE-ABS-KEY(*"topic model"*) OR TITLE-ABS-KEY(*"topic modeling"*) OR TITLE-ABS-KEY(*"topic modelling"*) OR TITLE-ABS-KEY(*"topic discovery"*) OR TITLE-ABS-KEY(*"discovering topics"*)) AND (PUBYEAR BEF *2022*)

[32] `https://www.ibm.com/about`

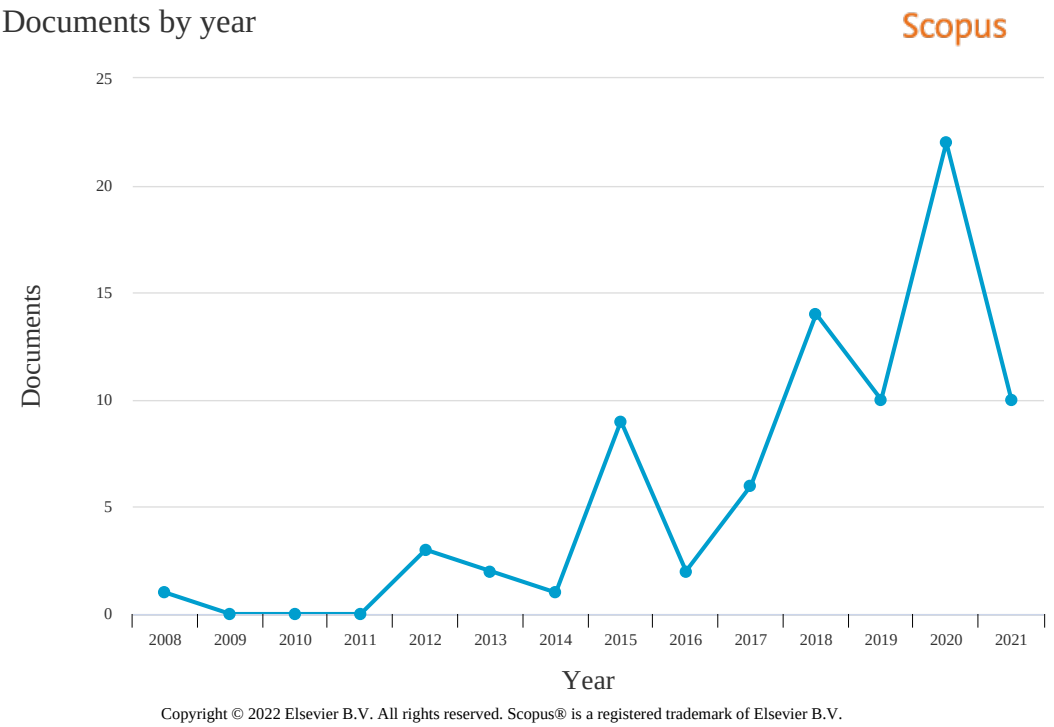


Figure 1.7: Research interest in topic modeling in Spanish and low-resource languages (documents by year). December 1, 2021.

[60] was performed several Machine Translation systems for indigenous languages of Latin America, where all systems presented was suffered from the minimal amount of data and the challenging orthographic, dialectal, and domain mismatches of the data. This shows us that there is still a long way to go in this field, so it is worth investing time and efforts in such systems.

According to the ALPAC (Automatic Language Processing Advisory Committee) report published in 1966 [61], there had been virtually no progress in the years prior to that report, most systems for solving any NLP task were modeled with lots of hand-written rules, which made the systems very expensive to develop and not at all flexible. At this time, Machine Learning-based models began to be developed, thanks to the increase in computational power and Noam Chomsky's (considered as 'the father of modern linguistics' [62]) ability to adapt to these types of problems. One of the first algorithms used was the decision trees, which allowed to creation of all that set of handwritten rules in an automatic way. From the 80s until a few years ago, there has been little development in artificial intelligence and NLP, being a stage in which no major advances have been obtained. In the 90s, n-gram or string of elements (more commonly words) became useful recognizing and tracking linguistic data sets, and numerical and statistical NLP methods had become remarkably valuable for the vast amount of online text. In 1997, recurrent neural network (RNN) LSTM (Long Short-Term Memory) models were introduced [63], but were not used for NLP tasks until 2007 [64].

The first neural 'language' model, a feed-forward neural network, was proposed in 2001 by Bengio et al. [65]. The advent of these Deep Learning-based systems put an end to this impasse, which, taking advantage of the great computational power we have today, made it possible to apply more efficient algorithms. As exposed, this progress also affected the field of natural language processing, making it possible to solve tasks that were so far almost impossible or that did not work as well as expected. Apple's Siri appeared in 2011 as one of

the first NLP/AI assistants for general consumers. Its automatic speech recognition module translated the owner's words into digitally interpreted concepts, where the voice command system matches those concepts to predefined commands, initiating specific actions [64, 66].

Currently, the research community is looking for text comprehension by the machine, and to achieve this, embeddings were born, a way of representing information in such a way that a machine can understand it. This type of data is what current models work with, so that they have the capacity, to a certain extent, to model the knowledge extracted from a text by means of a representation of a language unit. A specific type of embedding that appeared in 2013 is the word-to-word [67], known as word-embedding, which is a model that showed unprecedented results, capable of modeling a word, based on a vector of numbers. More particularly, if we apply vector operations we can obtain a modification of gender or verb tense, to convert our original word into another within a manipulable vector space (e.g., queen vs. king and swimming vs. swam). Sometime later, there was also appeared char-embeddings (character-by-character) [68] (2016), as well as sub-word-embeddings (2017), where a sub-word is an n-gram of characters [69], which usually resulted useful especially for low-resource settings.

In the same way, recurrent neural networks [70] (2010), specifically LSTM networks, which are able to process structured data sequentially, achieved remarkable performance in sentence and document modeling [71] (2013). These networks have the ability to work with temporal sequences, such as text, which is nothing more than a set of words (and characters) in a given order. This was a major breakthrough since the concepts would not be treated independently, but their order would be taken into account [72]. Similarly, with the work of Zhang et al. [73] in 2015, CNNs (Convolutional Neural Networks) proved to be useful for text comprehension, from character-level inputs to abstract concepts in the text.

Although these models performed quite well, they still had major problems, such as polysemy (i.e., when a word or linguistic sign has several meanings). In recent years, the concept of 'attention' [74] began to solve these issues, introducing Transformers models specific for NLP. Transformers are a series of models based on deep neural networks, and in particular attention-based ones [74] (2017). Attention is the influence that each word has within a text and there are different ways to consider it. This allows the model to decide which words or concepts to focus on so that the representation of the text will be much more accurate. It takes into account not only the words but also the structure and context, which solve, for example, the problem of polysemy.

The development of pre-trained systems such as BERT (Bidirectional Encoder Representations from Transformers) [12] (2018) and GPT (Generative Pre-trained Transformer) [75] (2019) are possible thanks to the additional training parallelization of transformers and allows training on larger datasets than was once possible. These systems can be fine-tuned for specific tasks or/and corpora [12, 76], despite were are trained with large language datasets, such as the Wikipedia[33] Corpus and Common Crawl.[34]

Thus, from the Transformers a multitude of variants were developed [77], generally obtaining state-of-the-art results over the traditional and contemporary ones. All of these models have marked a before and after in the industry and have begun to be used to solve a wide variety of problems.

1.2.5.2 NLP workflow

NLP tries to model language computationally and comprises two subtopics: language understanding (i.e., Natural Language Understanding - NLU) and language generation (i.e.,

[33] `https://www.wikipedia.org/`

[34] Common Crawl is an open repository of web crawl data that can be accessed and analyzed for free. From `https://commoncrawl.org/`.

Natural Language Generation - NLG). This dissertation focuses on text analysis with NLU, the typical text processing steps are as follows [78]:

- Preprocessing or text normalization: Text preprocessing is very important for most NLP problems and especially for 'raw text' coming from social media platforms, which is often 'noisier' than its counterpart in news or books. In social media, people often write informally with slang, emoticons, emojis, etc. There are several text normalization steps and their goal is to be able to use most of the information in the text. A key preprocessing step is *tokenization*, which consists of segmenting a text into tokens, usually words or sentences. Stop-words removal and spell checking are also useful in word normalization, as is lemmatization, which attempts to determine the lemma of a word (i.e., word at the head of an article in a dictionary or encyclopedia).[35] Like lemmatization, stemming is the process of deriving words down to their root form.

- Part-of-speech (POS) tagging: In POS tagging, given a sentence consisting of a sequence of words (tokens), POS tagging attempts to correctly 'tag' each word with its category or word class (verb, noun, pronoun, adjective, etc.), resulting in a large amount of information about a word and its adjacent words. POS tags can also be more fine-grained in providing information (as when they include morpho-syntactic information, such as gender, number, conjugation tense, form, etc.). In many tasks they are used to extract features, for example, to feed a model.

- Dependency parsing: Once the *morphological information* has been obtained, the dependency parsing represents the grammatical *syntactic structure* of the sentence, the relationships between the different elements, and the dependencies between them.

- Semantic analysis: Once the syntactic analysis of dependencies is finished, there follows a step to extract the semantics or meaning that the author of the text wants to communicate. That is, the *semantic analyzer* tries to discover the signification of each sentence.

- Pragmatic analysis: The *pragmatic analyzer* in a last stage, given that the sentences are related to each other, and it is these connections and references between them that construct the speech, aims to obtain the meaning of the this.

Figure 1.8 shows the six existing levels of linguistic analysis and structure [78]. The phonological levels are left out of this thesis, as we concentrate more on the outer layer levels.

1.2.5.3 NLP common tasks

The problems that NLP solves can be divided into various groups, commonly referred to as 'tasks'. Below, we will briefly conceptualize the most common ones, in addition to those already covered in this dissertation (i.e., sentiment analysis and topic model, see subsection 1.2.3 and subsection 1.2.4, respectively):

- Machine Translation: This task is one of the first tasks that were attempted to be solved, as mentioned above, in which one tries to convert one natural language into another. It is a task known as 'Sequence to Sequence' and usually traditional approaches try to solve it in such a way that each element of the input corresponds to its translation [79]. The input to the system is a sequence of tokens and an output of equal or different size is expected. Machine translation seeks a solution that is much more complex to obtain depending on the richness of each language pair [74].

[35] From `https://dictionary.cambridge.org/dictionary/english/lemma`.

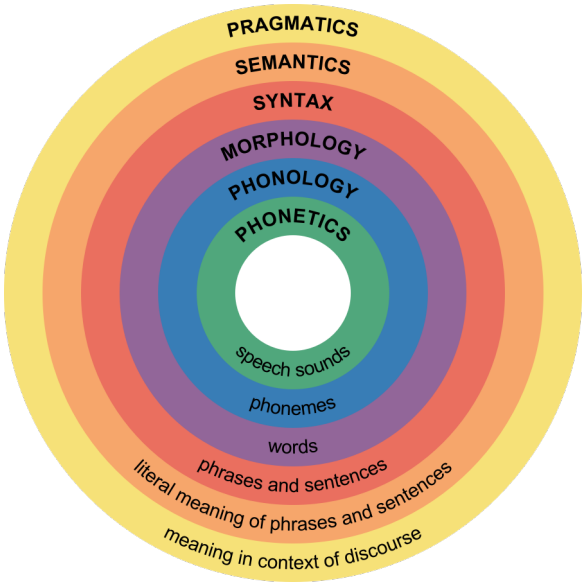


Figure 1.8: Major levels of linguistic structure [78, p. 110, Figure 4.1].

- Dialogue Systems: A dialog system is a system that interacts with the human being in natural language, which is developed mainly in text, but also in graphical, spoken and multimodal systems. A text-based dialog system is one in which you 'chat' with the system, while in spoken dialogue systems the spoken natural language interface performs the communication. On the other hand, multimodal systems are those that deal two or more combined modes of user input (e.g., speech, text, touch, manual gestures, etc.) in coordination with the output multimedia system [80].

- Information Retrieval: It is well known the vast amount of textual information present on Internet, which would not be accessible without information retrieval systems [81]. These systems are in charge of retrieving the relevant information that a person requires at a given time. It is a very active discipline,[36] due to the need to have reliable information available in the most immediate and efficient way possible to satisfy the user's demand.

- Information Extraction: This task consists of thoroughly exploring a text or message to extract those elements of information that are of interest. A large number of situations require applications such as these systems, which allow extracting entities in huge amounts of information [82, 83]. For example, in a news article is very important the identification of people, places and events, as well as related mentions.

- Named Entity Recognition: This is a word or token classification task and is very useful within NLP. It is related to *information extraction* and is based on finding entities within a text and determining what type they are, i.e., in a given text, words considered as entities are searched for and organized according to their type. Given a text, NER (Named Entity Recognition) detect the occurrence of names or expressions that refer to entities or other useful data. It is therefore widely used for parsing sentences, even in social media posts [84, 85].

- Text Generation: In NLG, the text generation is the process by which a model is able to create text, given a context. It has not been studied in depth until the advent

[36] Since 1971, with the first conference of the SIGIR (Special Interest Group on Information Retrieval). From `http://sigir.org/general-information/history/`.

of deep learning, where models have been able to generate robust and meaningful text [86]. It can be very useful for generating stories, writing text from key ideas or summarizing. Thus, these systems can take as input any source piece, from a random seed, another text or any sequence of properly processed data.

- Text Simplification: This task consists in the conversion of a complex text into a much simpler and easier to understand text. It is not the same as *Text Summarization*, since a summary is not the same as a simplified version of a document, but these may reinforce each other [87]. While a summary extracts the main idea of a text, a simplified text exposes the same complex text but arranged in a much more accessible way for understanding, so even it may sometimes require NLG to paraphrase the text [88].

- Question Answering: This task aims to give an answer to a question within a context, the context being a text where the answer to the question to be answered can be found. After understanding the question, the next step is to locate the information that answers the question, extract the exact answer and, finally, display the answer to the user. Usually the answer is a sentence extracted directly from the context, but there are some models that build it from scratch [89].

1.3 Hypothesis

It is common to assume that a text is written in a certain language, and even if that were true, each language has its own dialects that vary according to the area where it is spoken. So, it is also necessary to pay attention to this specific dialects. For instance, Spanish has various varieties such as Castilian, Mexican Spanish or 'Rio de la Plata' Spanish, among others. Based on this premise, we intend to study texts written in any language or the composed with a combination of two or more languages (even in a sentence or token level, known as code-switching), such as can be seen in opinions written with Anglicisms, or in Spanglish,[37] or in Jopara (Guarani-Spanish mixture). Therefore, the questions to be answered are the following:

- Can the techniques and methodologies developed for a particular language or dialect be applied, without losing performance efficiency, to other languages, even if two (or more) languages are mixed at the level of phrases and/or words?

- How to analyze these multilingual opinions?

Therefore, the relevance of the study we propose is given by:

1. Development of a line that has been little explored for truly low-resource languages (most of the efforts are focused on the English language, and without taking into account its varieties, dialects or idioms), in a growing field of study [90].

2. Proposal of approaches by adapting existing paradigms designed for multilingual opinion processing and analysis.

3. Development of resources (such as annotated datasets, linguistic corpora, machine-learning models, among others) for (i) topic discovery and (ii) multilingual sentiment analysis. Note that for (i), we focused on Spanish spoken in Spain, and in the case of (ii), on Guarani and Jopara.

[37] A mode of speech of some Hispanic groups in which lexical and grammatical elements of Spanish and English are mixed.

1.4 Justification

It should be noted that, unlike traditional machine learning models, those based on neural networks, and especially transformers, have high complexity, since they depend on a large number of parameters and different aspects, so they also require a large computational power. In addition, there are a plethora of models, based on the same ideas, but differing parts in their implementation or use and, of course, also in its performance in certain contexts and/or languages. For all these reasons, it is to be expected that each of them will have different behavior and more depending on the task to be worked on. Consequently, it is necessary to make a precise study of the behavior of these models, as well as of the traditional ones, in order to know which of them is better adapted to the problem of multilingualism and, above all, to the low-resource languages, i.e., under what conditions and in what situations.

It is difficult to find in the literature, so far, enough works for tasks involving multilingualism with truly low-resource languages (as is the case of Guarani, see chapter 5 or chapter 6) or a particular dialect (such as Spanish from Spain, see chapter 4). This is due to the fact that the development of this particular area was carried out thinking in languages of certain richness, such as a reasonable number of Wikipedia or Common Crawl pages, despite being considered as low-resource.

On the other hand, it should be noted that, since most of the NLP systems work in rich-resource languages, the work performed on this thesis may be useful both for the Spanish-speaking community (especially Spain) and the Guarani-speaking community (Paraguay and bordering countries such as Argentina, Bolivia, and Brazil). Therefore, there are many uses cases in different areas and disciplines that can benefit from the insights created by the approaches we presented in this work. For example, marketing, psychology, sociology, political, or other use cases in other domains, such as tourism or health informatics, and so on. Areas and disciplines in which it is necessary to be able to exploit all the opinions written in these languages in all their possible dimensions (sentiments, affections, type of language, etc.) and hence the importance of having sufficient and applicable resources to carry them out in the most democratic and accurate way possible.

1.5 Objectives

In this research, we take the direction of text mining and natural language processing, at the intersection between computation, artificial intelligence, and computational linguistics, with a clear emphasis on multilingualism in low-resource languages. The objective is to provide a better understanding of this problem (or related approaches) as well as a holistic analysis and view of it. Therefore, our proposed objectives are as follows:

1. Study the different existing machine-learning approaches, that aim the language-independence, especially the most current ones based on neural networks, with respect to multilingual opinions written in social media, even if was written with code-switching.

2. Develop new linguistic resources for text analysis in low-resource languages and dialects, especially the written in social media.

3. Build machine learning models for NLP of low-resource languages and dialects, especially in monolingual, multilingual, and code-switching settings.

1.6 Methods

In the following, we describe our methodology for creating machine learning systems for multilingual and low-resource language environments. We attempt to outline the common idea and workflow used in this thesis.

1.6.1 Machine Learning system workflow for Topic and Sentiment Analysis in multilingual opinions and low-resource languages

Our first methodology shown in Figure 1.9 and proposed in [91, 92], works well performing a sentiment analysis system for social media posts written in English (i.e., TripAdvisor restaurant reviews). However, this same methodological flow, which usually works for many machine learning systems for most languages (especially resource-rich ones), fails when trying to perform such systems on resource-poor and code-switching languages. Therefore, we decided to study a workflow that works for the thesis' case study, such as the one presented in the Figure 1.10.

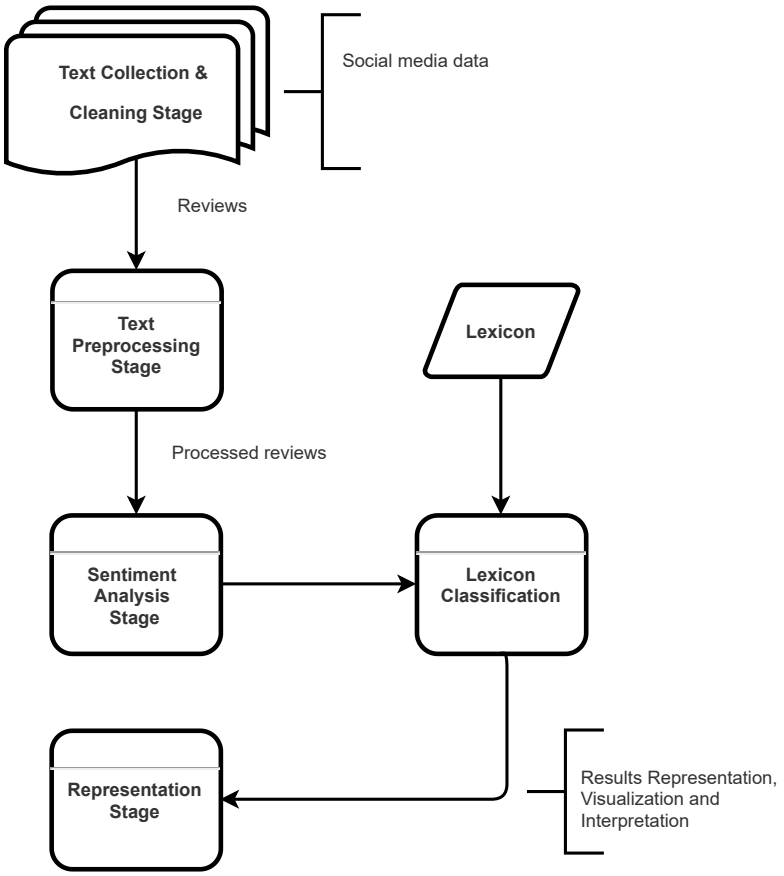


Figure 1.9: Methodology workflow of our `Gastro-miner` tool [91, p. 393, Figure 1 (left)].

In this way, we followed the data science life cycle paradigm, of course adapting it to work with low-resource languages on social media (see Figure 1.10): from the collection of opinions (from Twitter in our case, but could be relatively easily extended and adapted to other social media and text-content sources), their preprocessing and preparation (including their annotation in classes and/or labels), to their subsequent analysis, modeling, implementation, and representation.

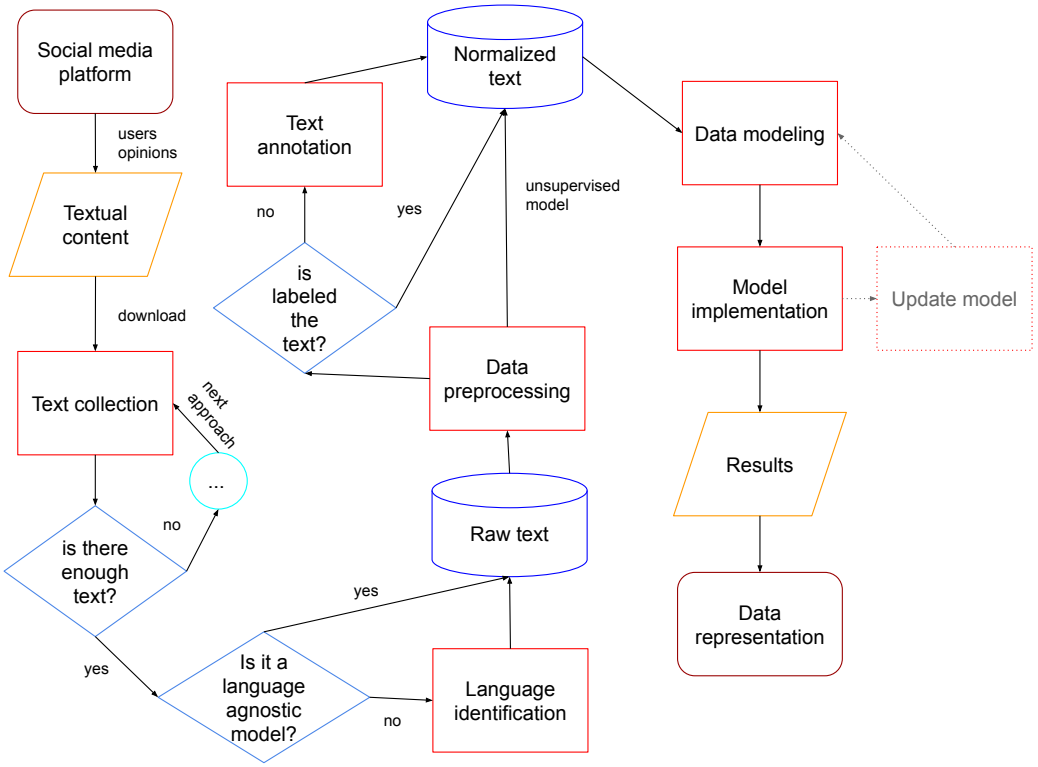


Figure 1.10: Data science life-cycle workflow adapted to textual content written in low-resource languages in social media platforms.

Workflow Figure 1.10 shows our data science life-cycle workflow adapted to textual content written in low-resource languages in social media platforms. The workflow starts with the social media platforms users' opinions, which are downloaded or collected under some approaches, e.g., keywords-based or specific-user-post-based. In this kind of process, particularly in low-resource languages is hard to find enough data, so it is necessary to try other approaches until it collects sufficient data for the model. If there are enough, according to if the model is language-agnostic or not, it needed to add an extra data preparation step, it is the case of the language identification or detection process. Note that, in low-resource languages settings even can be necessary to perform some extra effort to detect the language in the text due to the scarce of models that can be used to deal with this step. If the model does not need to perform language identification or the social media platform crawler allows us to know the language of the text, we can directly storage the raw data.

In the data preprocessing step (see again Figure 1.10), the text preprocessing to apply depends on the model to use, when traditional machine learning models, like Naive Bayes, require advanced linguistic features engineering, the more current ones, almost all of them neural-networks-based, require only some or nothing of data preparation, due to they are useful to deal with the feature engineering automatically. If the collected data allows us to use some metadata that can be useful as classes or labels, like the TripAdvisor's 'bubbles' (i.e., the user rankings that range from 1 to 5, useful for sentiment analysis for example), or if the model to implement pertains to an unsupervised type, like Topic Modeling, we can store immediately the normalized text to train the model. On the other hand, if we need to tag the text data, it is necessary to perform an additional process: the text annotation. In many cases, and especially when working with low-resource languages, it is necessary to include humans in the loop, because it is hard to find resources for these languages if it is desired to perform an automatic annotation with good quality. Note that, even human

annotators are hard to find for these languages, that resulting only in a few corpora with acceptable quality for this environment. In addition, usually, the annotation process is done by human experts of the data domain and preferably native speakers.

With the text normalized, we can perform the data modeling part (see again Figure 1.10), related to training, tuning, and validating machine learning models. In low-resource languages settings, generally approaches like cross-lingual or fine-tuning over pre-trained models, especially the multilingual and rich-resources ones, are preferred over training from scratch. Next to the data modeling process, the deploy and operational processes are required to implement the model. In these last steps, the model can learn and generalize, and be able to predict new and unseen samples. Note that, is usually the model is updated when is needed to add some extra feature engineering or when the quality of training data was improved, etc. This step is useful also in reinforcement-learning approaches. Finally, the data is represented to the system's final user across different visualizations, reports, etc.

Software and tools As for the tools used and developed in this thesis, they are mostly based on free and open-source software, mainly on a Python programming language's stack. All of them, implemented with a microservices architecture approach, under Kanban, a popular, flexible and continuos agile framework. A microservices architecture pattern decomposes a software system into a set of independently deployable services, according to the desired service's *business logic*, each with its own database [93]. For example, some tools used in this dissertation were:

- For data downloading and collection: Scrapy,[38] SNScrape,[39] APIs (Application Programming Interfaces), and so on.
- NLTK,[40] spaCy,[41] CoreNLP,[42] etc., for preprocessing, a customized tool[43] composed of several other tools (i.e., polyglot,[44] fastText,[45] langdetect,[46] langid,[47] and textcat)[48] for language identification and, Prodigy[49] for annotation.
- As well as lda,[50] scikit-learn,[51] TensorFlow,[52] PyTorch,[53] NCRFpp,[54] Hugging Face[55] and so on for analysis and modeling.
- Finally, for data visualization and representation we were used: Django,[56] Matplotlib,[57] Plotly,[58] among others.

[38] https://scrapy.org/
[39] https://github.com/JustAnotherArchivist/snscrape
[40] https://www.nltk.org/
[41] https://spacy.io/
[42] https://stanfordnlp.github.io/CoreNLP/
[43] https://github.com/mmaguero/lang_detection
[44] https://github.com/aboSamoor/polyglot
[45] https://github.com/facebookresearch/fastText/tree/master/python
[46] https://pypi.org/project/langdetect/
[47] https://github.com/saffsd/langid.py
[48] https://www.nltk.org/_modules/nltk/classify/textcat.html
[49] https://prodi.gy/
[50] https://lda.readthedocs.io/
[51] https://scikit-learn.org/
[52] https://www.tensorflow.org/
[53] https://pytorch.org/
[54] https://github.com/jiesutd/NCRFpp
[55] https://huggingface.co/
[56] https://www.djangoproject.com/
[57] https://matplotlib.org/
[58] https://plotly.com/

1.7 Structure of the thesis

Most of the content presented in this thesis is based on various research publications on the subject of this work.

Chapter 2 The aim of the chapter is to provide the community with an extensive review of work that has taken advantage of advances in deep learning to address the problem of multilingual sentiment analysis in social media. It provides a comprehensive overview of the field, identifying common ideas and problems that have been addressed in the implementation of multilingual SA. It also gives a reader-friendly summary and discussions to identify possible new research niches. This chapter is an extension of the paper published in the Special Issue 'Soft Computing for Recommender Systems and Sentiment Analysis' of the journal *Applied Soft Computing* [94] (see section A.4).

Chapter 3 The chapter is devoted to corpora creation for low-resourced languages and code-switched, and is divided into the following sections:

1. Collection of Spanish COVID-19 related tweets using keywords, language identification of the tweets (via several tools) and geolocation (via lists of Spanish cities and regions).

2. Guarani-Spanish (i.e., Jopara) Twitter text data collection effort for sentiment analysis, which dealing with certain limitations, like unbalanced classes due to the nature of the corpus, due to the scarce of tweets in Guarani-dominant.

3. Collection of three new corpora multiannotated of Jopara Guarani-dominant tweets for affect detection: (i) Emotion Recognition, (ii) Humor Detection (JOFUN) and (iii) Offensive and Toxic Language Identification.

The main content of the chapter has been extracted from the papers published at *Procesamiento Del Lenguaje Natural* [95] (see section A.3), *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* (co-located in *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*) [96] (see section A.5) and another work to be submitted to a journal.

Chapter 4 In this chapter, we automatically extracted topics to capture what Twitter users in Spain were discussing during the beginning of the COVID-19 pandemic. We used NLP strategies to perform an in-depth qualitative analysis of these tweets and the evolution of the topics (based on matching the extracted topics with newspaper news). We also proposed a small quantitative evaluation framework based on a human evaluation. To represent the topics, we used the traditional generative route, and also introduced a discriminative route, extracting the most salient keywords and phrases. The results of this work have been published at *Procesamiento Del Lenguaje Natural* [95] (see section A.3).

Chapter 5 In this chapter, several machine learning approaches (from traditional to more advanced, i.e. transformer-based [74]) are applied to a specific low-resource language: Guarani, as well as to its mixture with Spanish (i.e. Jopara). A brief discussion of the classification results provided by the different models was performed. In addition, an error analysis was carried out, which provided further information on the performance of the classifiers in this particular low-resource environment. The content of this chapter is an extension of the contribution published in the *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* [96] (see section A.5).

Chapter 6 The chapter summarizes our effort to build and pre-train several transformer-based language models [74] with Wikipedia data in Guarani, which is challenging due to Guarani is a low-resource language that also suffers from code-switching. We provide an affordable overview of the approaches we followed to train this set of BERT models [12] for Guarani and Jopara. We evaluated them on several tasks related to sentiment analysis, overall outperforming the existing BERT models, such as emotion recognition, humor detection, identification of offensive language and polarity classification. This chapter will be submitted to a journal.

Chapter 7 This part of the dissertation is a summary of the contributions and results obtained, which also lists the publications derived from our work.

Chapter 8 Finally, this chapter provides a summary of the conclusions drawn. Furthermore, we point out some open lines of future work derived from the results obtained.

Estructura de la tesis

La mayor parte del contenido presentado en esta tesis se basa en diversas publicaciones de investigación sobre la temática de este trabajo.

Capítulo 2 El objetivo del capítulo es proporcionar a la comunidad una amplia revisión de los trabajos que han aprovechado los avances del aprendizaje profundo para abordar el problema del análisis de sentimientos multilingüe en los medios sociales. Proporciona una visión general del campo, identificando las ideas y los problemas comunes que se han abordado en la aplicación del análisis de sentimientos multilingüe. Además, ofrece un resumen de fácil lectura y debates para identificar posibles nuevos nichos de investigación. Este capítulo es una extensión del artículo publicado en el número especial denominado 'Soft Computing for Recommender Systems and Sentiment Analysis' de la revista *Applied Soft Computing* [94] (ver §A.4 para más detalles).

Capítulo 3 El capítulo está dedicado a la creación de corpus para las lenguas consideradas como *low-resource* y con *code-switching*, y se divide en las siguientes secciones:

1. Recogida de tuits relacionados con COVID-19 en España, mediante palabras clave, identificación de idioma de los tuits (a través de varias herramientas) y geolocalización (mediante listas de ciudades y regiones españolas).

2. Esfuerzo de recopilación de datos de texto de Twitter en guaraní-español (es decir, jopará) para el análisis de sentimiento, lidiando con ciertas limitaciones, como clases desequilibradas debido a la naturaleza del corpus (debido a la escasez de tweets en guaraní-dominante).

3. Recopilación de tres nuevos corpus multianotados de tweets en jopará, con guaraní-dominante, para la detección de afectos: (i) Reconocimiento y detección de emociones, (ii) Detección de humor (JOFUN) y (iii) Identificación de lenguaje ofensivo y tóxico.

El contenido principal de este capítulo ha sido extraído de los trabajos publicados en *Procesamiento Del Lenguaje Natural* [95] (véase §A.3), *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* (co-ubicado en *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*) [96] (ver §A.5) y otro trabajo que se presentará a una revista.

Capítulo 4 En este capítulo extrajimos automáticamente temas para capturar lo que los usuarios de Twitter en España estaban discutiendo durante el comienzo de la pandemia de COVID-19. Utilizamos estrategias de PLN para realizar un análisis cualitativo en profundidad de estos tweets y de la evolución de los temas (basado en el cotejo de los temas extraídos con las noticias de los periódicos). También propusimos un pequeño *framework* de evaluación cuantitativa basado en una evaluación humana. Para representar los temas, utilizamos la ruta generativa tradicional, y también introdujimos una ruta discriminativa, extrayendo las palabras y frases más destacadas. Los resultados de este trabajo se han publicado en *Procesamiento Del Lenguaje Natural* [95] (véase §A.3).

Capítulo 5 En este capítulo se aplican varios enfoques de aprendizaje automático (desde los tradicionales hasta los más avanzados, es decir, basados en *transformers* [74]) a una lengua *low-resource* específica: El guaraní, así como a su mezcla con el español (es decir, el jopará). Se realizó una breve discusión de los resultados de clasificación proporcionados por los diferentes modelos. Además, se llevó a cabo un análisis de errores, que proporcionó más información referente al rendimiento de los clasificadores en este entorno *low-resource*

tan particular. El contenido de este capítulo es una extensión de la contribución publicada en el *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* [96] (ver §A.5).

Capítulo 6 El capítulo resume nuestro esfuerzo por construir y preentrenar varios modelos de lenguaje basados en *transformers* con datos de Wikipedia en guaraní, lo cual es un reto debido a que el guaraní es una lengua *low-resource* que también sufre de *code-switching*. Proporcionamos un resumen amigable de los enfoques que seguimos para entrenar este conjunto de modelos BERT [12] para el guaraní y el jopará. Los evaluamos en varias tareas relacionadas con el análisis de sentimientos, superando en general a los modelos BERT existentes, como el reconocimiento de emociones, la detección de humor, la identificación del lenguaje ofensivo y la clasificación de la polaridad. Este capíulo se enviará a una revista en la brevedad posible.

Capítulo 7 Esta parte de la tesis es un resumen de las aportaciones y resultados obtenidos, que también lista las publicaciones derivadas de nuestro trabajo.

Capítulo 8 Por último, en este capítulo se presenta un resumen de las conclusiones obtenidas. Además, se señalan algunas líneas abiertas de trabajo futuro derivadas de los resultados de nuestra investigación.

Chapter 2

Deep Learning Approaches for Multilingual Sentiment Analysis on Social Media: From State of the Art to Future Directions [1]

This chapter provides an overview of work that has taken advantage of advances in deep learning to address the problem of multilingual sentiment analysis in social media by following a language-agnostic way. It identifies the common ideas and problems that have been addressed, while providing an easy-to-read summary and thought discussion, as well as suggesting possible new research niches.

2.1 Introduction

Sentiment Analysis (SA) allows to automatically assess people's opinions towards products, services, and other entities. This knowledge can help make better decisions by seeking to improve key performance indicators. In addition, the massive adoption of social media such as Facebook and Twitter, e-commerce platforms and services such as Amazon, and even specialized review sites such as Rotten Tomatoes, have unleashed a wealth of content that needs to be analyzed. This data is naturally multilingual and multicultural, so analysis based on a single language can run the risk of not capturing the big picture [97]. In addition, there are significant challenges that may prevent the full exploitation of this data. Except in a few cases, such as English, most languages lack well-maintained and widely used resources for SA, such as annotated corpora and lexicons. Secondly, it may not be straightforward to adapt the same SA model to different languages, for example, due to variations in word order or usage, or noise introduced by machine translation. In addition, we have code-switching content, where users express their opinions using a mixture of languages in the same sentence.

Multilingual Sentiment Analysis (MSA) is an attempt to address these issues through various strategies. For example, in multilingual sentiment analysis, it is characteristic to take advantage of resource-rich languages to perform SA in a resource-poor language. In addition, the development of language-independent models that can handle SA in different languages or a code-switching configuration.

There is a wide spectrum of approaches to SA, e.g., [15, 98, 99, 100], which can be based

[1] Chapter based on a small extension of the content of a paper published in the Special Issue 'Soft Computing for Recommender Systems and Sentiment Analysis' of the journal *Applied Soft Computing* [94] (see section A.4).

on supervised methods but also on unsupervised methods, exploiting sentiment lexicons, grammatical parsing and syntactic patterns. In section 2.2.1 and 2.2.3, we include an overview of the different formulations of this task as well as the evolution of SA and MSA. More recently, deep learning (DL) approaches have become a trend leading to state-of-the-art results, with authors such as [101, 102, 103] exploring Convolutional Neural Networks, Adversarial Networks and Recurrent Neural Networks among other models. In section 2.2.2 we summarise some of the advances in deep learning for SA as an introduction for the main topic of this chapter, the applications of deep learning in multilingual sentiment analysis in social media.

Using the methodology detailed in section 2.3 as a guideline, we curated and reviewed 24 relevant research papers. We categorized them as regard the main idea in multilingual, cross-lingual or code-switching approaches, covered in section 2.4.1, 2.4.2 and 2.4.3 respectively. For each one, we discuss its distinctive contributions, the experimental setup, corpus, and main results. Also, section 2.4.4 includes a comparative that allows a quick and broad view of the advances in the domain. This analysis drew interesting conclusions such as the few works, to date, leveraging recent developments in contextual embedding. Other main findings and conclusions are covered in sections 2.5 and 2.6.

As sentiment analysis and in particular, deep learning approaches has been growing as an important research field, there have been early efforts [48, 97, 104, 105, 106, 107] to systematize the knowledge corpus in this domain, works extended lately by [108, 109, 110]. Recently, [111] examined the fundamentals of the multilingual case. However, more than seventeen works we identified introducing or exploring specific ideas for MSA have not been studied by the aforementioned works. Another of our main contributions is to drive the review by the underlying hypothesis of each work, not only analyzing them as regards the type of neural network they used. This is important since the same task can be tackled by very different ideas. Also, we focused the analysis on the current three major strategies for MSA: multilingual, cross-lingual, and code-switching. This high-level view of the domain can help to unveil interesting patterns more than the type of neural network implemented. For example, the use of adversarial training to learn language-agnostic features.

2.2 Background

In what follows, we briefly review sentiment analysis and multilingual sentiment analysis as well as deep learning applications for sentiment analysis.

2.2.1 Sentiment analysis on social media

Starting from Wiebe et al. work [46] in the late 90s, there has been a surge of interest in the different setups of SA. In general, it can be done at a document, sentence, or aspect level [100] and the classification in terms of positive, negative, or neutral, but also other more fine-grained scales such as a ranking from 1 to 5. This attention over SA is closely tied to Social Media in its key role in the rise of modern SA particularly with the works of Pang et al. [112], and Turney [47], in 2002. The first used machine learning (ML) classification techniques over movie review data (outperformed human-produced baselines), and the second, achieved an average accuracy of 74% for his recommendations based on online reviews of automobiles, banks, movies, and travel destinations, which used Semantic Orientation (SO) applied to unsupervised classification. Later, Pang and Lee (2008) [98], focused on the fundamentals and basic applications of SA, with a list of resources such as lexicons or datasets.

A comprehensive review that shows the maturity of SA up to 2012 can be found in the book of Liu [15]. This work covers most of the topics, definitions, research problems

(e.g. opinion spam detection), types of opinions (such as explicit and implicit opinions), and classification algorithms for SA. In 2013 Feldman [113] and Cambria et al. [114] wrote about the basic techniques, key tasks and applications as well about the evolution of the field.

Another source to take the pulse of the continuous advances in the field has been the tasks related to SA in Twitter hosted by the International Workshop on Semantic Evaluation (SemEval) from 2013 to 2017 and in 2020 for code-switching text. From the latest results, we can corroborate a shift toward the application of deep learning with 20 out of 48 systems participating in SemEval 2017 [16]. In the next section, we overview some of the recent advances in DL applied to SA without considering the multilingual task.

2.2.2 Deep learning on sentiment analysis task

Deep Sentiment Analysis (DSA) relies on the great potentials of DL shown for NLP tasks. Here, we briefly commented on some examples to illustrate how DL has been leveraged within the SA.

Word embeddings are used for language modeling and feature learning. They are commonly used as an input of the DL models, being Word2Vec [115] and GloVe [116] two frequently used approaches. Also, there are contextualized embeddings such as ELMo [117], which represent better the polysemy of the words. Besides using pre-trained embeddings, they can be learned to encode some specific task semantics. In the context of SA, this approach has been explored in works such as [118] and [119].

Another trending field within DL is the attention mechanism [74], which allows the model to non-uniformly weigh the contribution of the context when computing the output. This is another choice that is being used frequently in SA, for example, to capture the interaction between aspects and its context as in [120, 121].

Also, there has been a great interest in novelty architectures for SA. One approach that has been received considerable attention, in particular when working at the document level, is the design of hierarchical models that learn a representation for sentences from its words, on top of this level, another model can learn representations for documents. Different alternatives such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) can be used at each level. Works in [122], [123] and [124] are some examples of this approach.

As of last additional example of the ideas that have been explored, not only for monolingual SA but also for MSA we mention the use of adversarial learning to produce a set of domain-independent features. This is the hypothesis of works such as [125] and [126] for cross-domain SA.

The concepts discussed in this section do not exhaust the applications of DL to SA, a more complete revision can be found in [127].

2.2.3 Multilingual sentiment analysis

In this section, we introduce the fundamentals of multilingual sentiment analysis as well as some of the earlier applications of DL to MSA. Initially, the applications of SA have been developed basically for one language, English in most cases, but the multilingual nature of Social Media has shifted the field to a multilingual analysis. Also, advances in SA backed by DL have made it possible to include low-resource languages and avoid the use of translation tools.

A frequent approach for MSA is called Cross-Language (also Cross-Lingual) Sentiment Classification [128] which relies on machine translation [129] [130]. For example, [131] reported an improvement over non-DL classifiers (SVM - Support Vector Machine) translating

from other languages (Hindi, Marathi, Russian, Dutch, French, German, Portuguese, Spanish, and Italian) to English, to use augmented word embeddings together a CNN model. However, the cross-language approach carries several issues and weaknesses, for example, notable discrepancies in the data distribution, potential cultural distances even in a perfect translation, hard and costly translation tasks for large corpora with issues as charges, availability, performance [128].

Another usual setup for SA is the code-switching one, also called code-mixing or code-mixed. In this case, the content to be analyzed is expressed alternating two or more languages even in a single sentence. One early approach that takes on this problem with DL is Wang et. al [132] for Chinese-English. They proposed a bilingual attention LSTM to perform SA in a corpus from Weibo.com capturing the informative words from both the bilingual and monolingual contexts. Code-mixing is frequent in social media sites such as Facebook and Twitter in countries with a large part of its population speaking more than one language, for example in India where several official and non-official languages are used. This is an active research area, in particular for Hindi-English. An early proposal [133] is based in MLP (Multi-Layer Perceptron) with word-level features for Hindi-English and Bengali-English Facebook posts. A more detailed summary of SA for Indian languages with a special focus on the code-mixed text can be found in [134].

There has been a lot of interest to systematize the advances in MSA. For example, [135] and [136], the latter reviewed the principal directions of research focusing on the development of resources and tools for multilingual subjectivity and SA and addressed both multilingual and cross-lingual methods. Singhal & Bhattacharyya (2016) [106] described some of the different approaches used in SA research. Lo et. al (2017) [97] revised various of the main approaches and tools used for MSA at the time. They identified challenges and provided several recommendations with a framework for dealing with scarce resource languages. Also in [48] and [137] we can find revisions of the field, however, they did not delve into the use of DL in MSA.

2.3 Methodology

This chapter aims to review the research in multilingual sentiment analysis on Social Media with an emphasis on those featuring deep learning approaches, most framed within the last four years, which led us to examine the current approaches, methodologies, or tools used in Deep Multilingual Sentiment Analysis (DMSA).

Journal articles and conference proceedings that met the following search criteria were included: had to investigate (a) deep learning AND (b) sentiment analysis AND (c.1) multilingual OR multi-language OR multilanguage; (c.2) crosslingual OR cross-lingual OR cross-language OR crosslanguage; (c.3) code-mixed OR codemixed OR code-mixing OR codemixing OR code-switching OR codeswitching; (c.4) bilingual OR bi-lingual. Moreover, we searched in well-known databases such as Scopus and Web of Science. The data were extracted by one author and checked by another.

The search included studies published between January 2017 and December 2020, that because the shift toward the application of DL-based sentiment analysis happened in 2017 as shown in [16]. Our last search took place in January 2021.

The search yielded 96 studies that were examined to ensure they satisfy the following criteria:

a) must handle explicitly the multilingualism either by

   a.1) training with one or more languages and evaluating with a different one or others,

   a.2) train with a multilingual corpus, i.e., the same model sees text in different languages during training regardless of this being at different steps.

Thus, we excluded works that separately trained and evaluated the same architecture in different languages, i.e., created one model for each language trained only with data from the given language. We check the candidates for elimination.

We also revised the citations from the selected studies as well those referenced by previous reviews [44, 97, 107, 108, 109, 110, 111, 137, 138, 139] to identify possible candidates, applying the filters (a.1, a.2). In total, 24 publications were eligible for review.

From the chronologically incorporated researches, we extracted data and information about (i) research characteristics, as authors, year of publication, languages covered, methodology, corpus characteristics; (ii) sentiment level and categorization (binary, ternary, or fine-grained, e.g., rates 0-5); (iii) deep learning architectures and techniques, and (iv) results and effectiveness of the proposal against baselines or state-of-the-art models. We also reviewed each work to identify the underlying idea to handle multilingualism, with a focus on the results that assessed the hypothesis.

2.4 Deep learning techniques for multilingual SA on social media

In this section, we review a large corpus of research related to the applications of deep learning to multilingual sentiment analysis. Instead of driving the analysis by the type of architecture or techniques, we choose to organize the works by its sub-domain within MSA, i.e. multilingual, cross-lingual, or code-switching approaches. Inward each category, we proceeded chronologically to track the evolution of the field but separated the aspect-based studies since in general, they lead to very specific architectures. Also, we aim to provide a high-level perspective considering the underlying hypothesis of each work. Finally, the epigraph 2.4.4 aids the reader to take a glimpse of the domain as regards the models, baselines, corpus, core ideas, and languages covered. For clarity, and due to the variety of datasets and languages covered by each work, Table 2.1 illustrates the corpus and reports the number of tweets/sentences/documents used.

2.4.1 Multilingual approaches

This category groups a large set of works that aims to be language-agnostic to those seen during the training. Common goals are the design of systems capable to learn directly from multilingual unpaired content and providing predictions regardless of the source language.

2.4.1.1 Sentence-based studies

Training the same model for different languages is explored by[140]. They fit a multilayered CNN in two phases. First, they learn word embeddings from a large corpus of $300M^2$ unlabeled tweets in English, Italian, French, and German. The parameters are optimized further during the second stage, trying to infer weak labels inferred from emoticons. Finally, they fine-tuned the model using a corpus of annotated tweets. Experiments evaluated a model (FML-CNN) trained in all languages at once, a model fitted in a single language (SL-CNN), and other variations. The results showed that FML-CNN reaches slightly worse performance, about 2.45% lower F1 score, compared to SL-CNN (67.79% for Italian). However, experiments suggest that FML-CNN can handle better for code-mixed text.

Another hypothesis is to exploit character-level embeddings to achieve language independence. In [128] and [141] authors describe language-agnostic translation-free architectures (Conv-Char-S, Conv-Char-R) for Twitter-based on a CNN that can be trained in several languages at once. They evaluated their approach using tweets in English, Portuguese,

[2]M: Million.

Spanish, and German from the corpus in [142] achieving an F1 score above 72.2% [141] for the multilanguage setup. The slightly worst results with respect to some baselines such as LSTM-Emb [143] can be a trade-off since the models have ≈ 90 times fewer parameters and use ≈ 4 times less memory.

The idea of multilanguage character embeddings is explored also by [144] but mapping of each character to its UTF-8 integer code. The architecture (UniCNN) is similar to [128, 141], placing a CNN after the embedding layer, with a fully connected classifier at the top. They used a subset of the Twitter corpus in [142] and the *CrisisLexT26* [145] corpus [3]. The UniCNN achieved accuracy $\geq 75.45\%$ for all languages. Moreover, except for English, they outperformed models that require translation or/and tokenization such as TransCNN (Word), a similar architecture that operates at word level and translated text (79.57% for English).

In [146] the same authors further developed the concept but within an architecture (Word-Character CNN) that processes the text through two parallel CNN, one for words and the other for characters. The hypothesis is that words and character features provide complementary information. Outputs from both CNN are merged before being feed to a fully connected classifier. To achieve language-independence, the embedding layer is kept trainable. They used the same Twitter corpus as in [144]. The hybrid model yields a better performance ($\geq 77.13\%$) compared to pure word/character CNN such as [103] ($\geq 74.64\%$), [128] ($\geq 75.41\%$) and their former model UniCNNs [144] ($\geq 75.45\%$) for languages already studied in [144]. Its interesting that the two romance languages considered, Spanish (69.82%) and Portuguese (72.87%), had the worst performance.

Medrouk & Pappa [147] studied a similar architecture. It comprises a stack of CNN working as a feature extractor, i.e., an encoder, followed by polling and a fully-connected predictor, but in this case, working at the n-gram level. To this point, CNN seems a popular choice within the domain in contrast to LSTM. The model is fed with reviews written in French, English, and Greek without any language indication. Empirical evaluation over a mix of contents from three languages yielded an F1 score of 88%. These results reinforce the assumption that the n-gram CNN can produce language-independent features capturing the local relations between words useful for multilingual polarity analysis.

In [148] the same authors investigated whether CNN and LSTM variants of the embedding-feature extractor-classifier architecture need extra pre-training hassle or additional complexity to handle multilingual data. Experiments were conducted training monolingual and multilingual models, achieving accuracy over 90% for both types of networks working at the n-gram level. Moreover, the fact that the multilingual models behave as well as the monolingual ones, seems to confirm the hypothesis about their ability to extract rich features without distinction if processing single or multilingual datasets.

2.4.1.2 Aspect-based studies

Regardless of the promising results for SA at sentence level that achieves simpler architectures such as [148], it is not a surprise that for aspect level authors proposed more complex models.

The architecture (GRCNN-HBLSTM) proposed by [149] combines two word-level feature extractors. A BiLSTM encodes sentences that take as inputs the embeddings for the topics, the aspects, and the words. The original word embedding and a feature vector from a character CNN are combined through a gate mechanism to achieve language independence. The second encoder is a regional CNN that aims to preserve the temporal relationship between different sentences, also capturing some of the long-distance dependencies of the aspects. Both feature sets are feed to a sentence-level BiLSTM together with an attention

[3] tweets related to 26 crisis events that happened between 2012 and 2013 in more than 40 languages

mechanism. A softmax classifier handles the output of the last layer. In experiments using a subset of the dataset in [150] their full model yields an F1 score above 78.04% in all cases outperforming baselines such as the Hierarchical LSTM [122] ($\geq$ 78.04%). What is more important, they compared a version (CNN-HBLSTM) without the gate mechanism that achieves the worst results ($\geq$ 74.66%) and highest variance among languages.

Summary So far, we have reviewed the purely multilingual approaches for SA. We have contrasted very different approaches. However, at the sentence level, the common strategy is to learn features from a multilanguage set using CNN and feed a classifier module with those features. Unsurprisingly, for the aspect SA setup, authors embrace more complex architectures leveraging attention mechanisms and aspect embeddings. The next section is devoted to the cross-lingual category.

2.4.2 Cross-lingual approaches

We categorized as cross-lingual the proposals where the focus is to leverage resource-rich language assets to extrapolate to a low-resource target, for example, through transfer learning. This is the core of the proposal in [151] for the cross-lingual projection of sentiment embeddings. Their custom architecture (Dual-Channel CNN) has one channel which works with word embeddings to extract features through convolutions. The other channel is similar but uses word sentiment embeddings which can boost the classification. Features computed from each channel are merged before being feed to a fully connected layer. It is worth noting that for low-resource languages, the sentiment embeddings can be projected from English. They evaluated their approach for English as the source and Spanish, Dutch, German, Russian, Italian, Czech, Japanese, and French. The induced embeddings lead to better results in 7 out of the 10 trials with accuracy over 79.30%.

While not common in cross-language SA, in [152] authors explored the architecture comprised of a feature extractor (BiLSTM) feed by embeddings followed by the classifier (dense layer) for transfer learning. First, they use a large dataset to train the whole model. Afterward, they fine-tuned in a small labeled dataset from the target language, but only the embedding layer remains trainable. They trained using TripAdvisor reviews in English for the first stage and tweets in Greek for the second, being the results very sensible to the size of this dataset (accuracy drops from 91.7% to 73.2% as the dataset shrinks from 400 to 330). As the authors note, it will be interesting to study how a different degree of syntactic similarity between languages influences results.

Next works reviewed within the cross-lingual category explored the idea of using adversarial training to learn a set of language-independent features.

In [153] the authors delve into the synergies of microblog data in different languages from the same user to extract personalized language-specific or independent features to alleviate the lack of data in some sources. The architecture has four components. First, an attention mechanism encodes users as feature vectors to propagate their individuality across the system. The second component are encoders θ^1, θ^2, one for each language, computing specific-language features. The third element is the language-independent encoder θ^G that is fed with sentences from both languages as well the user attention vector. Encoders are CNN over different n-gram representations of the sentences and an attention mechanism for the user-specific information. The classifier is softmax layer for each language with inputs from θ^G and θ^1 or θ^2. The last module is a Generative Adversarial Network (GA) that drives θ^G to a set of features useful for SA when combined with θ^1 or θ^2 but uncorrelated with the language of the input sentence. Experiments with Twitter and Sina Weibo compared monolingual baseline models and the proposal, trained with both languages at once. Results show an increase of the F1 score up to 2.12% respect the best monolingual ($\geq$ 79.85%).

Another work that leverages adversarial training to learn a set of language-independent but highly discriminatory features is [154]. The architecture (ADAN) uses the Deep Averaging Network (DAN) in [155] as a feature extractor with a Bilingual Word Embedding (BWE) [156] as an input layer. These features are feed to the classifier and to a language discriminator acting as an adversarial driving DAN to language-independent features. Empirical evaluation shows ADAN (accuracy $\geq$ 42.49%) improves in at least 6% over a version without the adversarial module (only DAN) trained with English to predict Chinese and Arab. It suggests that the adversarial mechanism is crucial for the results.

The adversarial mechanism is also critical in [157]. They build a cross-lingual word embedding using an Adversarial Auto Encoder (AAE) [158] feed from the outputs of two LSTM, one for each of the source and the target languages. On the top of this module sits a classifier based in Bidirectional Gated Recurrent Unit (BiGRU). Evaluating using Amazon comments with English as the source and Chinese and German as the target, the model (TL-AAE-BIGRU) achieved an F1 score $\geq$ 78.13%, about 3.44% better as average than the model without AEE ($\geq$ 73.25%). This result is consistent with [153, 154] who noticed the benefits of the adversarial module too.

Instead of adversarial training, in [159] they opted to create universal embeddings by the combination of embeddings from high and low resources languages. The Universal SA (UniSent), involves the pre-training of two BiLSTM on English (labeled tweets), the alignment of low-resource language embeddings to the English ones with an unsupervised and domain-adversarial approach (MUSE [160]), and the fine-tuning of the low-resource languages validation sets applying an *Universal Embedding Layer*. This layer represents a word in a low-resource language by the weighted average of the most similar words to it in the English word embedding. The embeddings are feed to the classifier, a many-to-one LSTM layer. The experiments were carried in texts from OpeNER for Spanish and MultiBooked for Catalan. UniSent achieved F1 score $\geq$ 81.4% for the binary classification and $\geq$ 54.2%, outperforming even a version tested using translated texts ($\geq$ 74.0% and $\geq$ 40.6%).

In [161] authors favor the more simple architecture comprising LASER (Language-Agnostic SEntence Representations) toolkit[4] as a language-independent embedding, a feature extractor (CNN or BiLSTM) and the classifier. The multilingual embedding aims to drive the model to language-agnostic representations, being able to perform SA in languages different from the ones seen during training. Experiments training with Polish (F1 score 79.91%) to predict other languages seems to evidence this premise since in all cases, F1 was $\geq$ 77.96%. Also, that the setup LASER+BiLSTM is better in general.

A similar architecture is proposed by [162] but using Bilingual Bag-of-Words without Word Alignments (BilBOWA) and VecMap[5] as cross-lingual embedding. Experiments over English and Persian electronic product reviews evaluated different alternatives for the embeddings and the feature extractors (CNN, LSTM, CNN-LSTM, and LSTM-CNN). The VecMap+LSTM-CNN with dynamic embeddings, i.e., fine-tuning the embeddings with the training data, achieved the best results with an F1 score of 91.82%.

Summary So far, we have reviewed cross-lingual approaches for SMA. Though there are very different perspectives to solve the problem, they also shared some common ideas, as the projection of resources such as sentiment embeddings or from a resource-rich to a resource-poor language.The next section is committed to reviewing the works that addressed the code-switching setup.

[4] `https://github.com/facebookresearch/LASER`
[5] `https://github.com/artetxem/vecmap`

2.4.3 Code-switching approaches

In this section, we examine the reports that tackled code-switching sentiment analysis. This setup poses challenges such as spelling variations, transliteration, informal grammar forms, and the scarcity of annotated data.

The underlying idea of [163] is to learn a sentiment feature space preserving the similarity of sentences in terms of the sentiment they convey. This enables us to measure the relatedness between code-switched content and labeled data from a resource-rich corpus. The Sentiment Analysis of Code-Mixed Text (SACMT) architecture uses a siamese BiLSTM with tri-gram embeddings as input and a fully connected layer on the top. They compared a model trained only with pairs of code-mixed text (Hindi-English) with an F1 score of 67.2%, to other trained with pairs of sentences one from English and other code-mixed (75.9%). These results suggest that in effect, the model benefits from the additional data provided by the English corpus.

Most of the research in the code-mixed text has focused on the English-Hindi setup. One exception is the seminal work on code-mixed Bambara-French Facebook comments in [164]. They examined different variations of the base architecture with embeddings (at word or character level) as input, followed by a feature extractor (one of LSTM, BiLSTM, CNN, CNN-LSTM) and finally, the classifier. To mitigate the lack of pre-trained embeddings in Bambara, the model learns multilanguage embeddings from characters or words in the code-mixed corpus. The best performing model was a one-layer CNN model with an accuracy of 83.23%. The comparison between LSTM and CNN as feature extractors, where the latter-one yields better results, is coherent with a noticeable preference for this type of model within the domain.

One problem when working with code-mixed data is the noise and the small size of datasets. To alleviate this, in [165] authors proposes to use n-gram embeddings instead of the sub-word ones suggested by [166]. Another novelty idea within the domain is the model that works as an ensemble of a Multinomial Naïve Bayes (MNB) and a recurrent neural network (LSTM or BiLSTM) classifiers. They trained the MNB using both word-based 1 and 2-gram features while the neural networks models with the character 3-grams. The results, together with those reported by [166], suggest that for the LSTM network, the 3-grams representation is a better option (F1 score of 58.6%) over characters (51.1%). The sub-word level encoding achieved 65.2% but, the difference between the architectures can mislead the conclusions. Values for the ensemble (66.1%) show that this can be of benefit.

Authors of [167] assessed different versions of the architecture that combines sequentially one feature extractor and a classifier. The first one, a document to vector (Doc2Vec) layer, whose output fed an SVM. The other three featured a FastText classifier, a BiLSTM, and a CNN with a softmax. The last two had a trainable word embedding layer before the NN. They curated a new corpus for Kannada-English (achieving best results with the CNN model with an accuracy of 71.5%). Also, they evaluated using two available Hindi-English and, Bengali-English corpus [168] with the BiLSTM model achieving slightly better results, 60.22% and 72.2% respectively.

In [169] authors delve into whether to use characters, words, or sub-words. Also if projecting code-mixed text to a single feature space is a rich-enough representation for SA. Their architecture (CMSA) combines three parallel feature encoders before a classifier of four dense layers. The collective encoder aims to represent the overall sentiment of the sentences. It is based on a BiLSTM network whose end states are the features. The specific encoder is also a BiLSTM, but in this case, the intermediate states are also considered features through an attention mechanism. Both encoders take as input the output of a sub-word level CNN. The last one is a set of hand-crafted features to augment the information supplied to the model. They evaluated the effect of removing some of the components. CMSA achieved an F1 score of 82.7%, better than the model only with the specific encoder (80.1%), or only the

collective (79.5%). It seems to support the hypothesis about the synergies of the different representations.

Another model combining different feature extractors is the proposed in [170]. Similar to [146], they have character-level and word-level feature encoders. The first one is inspired in [166], stacking a CNN followed by two LSTM layers. It aims to help with language independence, noise, and non-standard spelling. The other extractor stacks two LSTM layers. The concatenation of the two feature sets is the input for the classifier, a stack of two dense layers, and a softmax. Their model achieved an F1 score of 66.13% over the Hindi-English corpus in [166] improving about 5.5% the baseline in [166]. As in [146], combining both types of feature extractors seems to lead to better results.

Recently in [171] authors evaluated several architectures (BiLSTM-CNN, DoubleBiLSTM, Attention-based), each one with and without GloVe, and BERT (Bidirectional Encoder Representations from Transformers) [12] - over tweets and Facebook comments for English-Hindi and English-Bengali. The same models were trained in a monolingual corpus to observe the effects of the code-mixing. The best model for code-mixing, the Attention-based model with custom word embeddings, achieved an F1 score average of 0.66 and 0.67 for the monolingual setup. It is interesting that the performance of $\text{BERT}_{base-uncased}$ (English), the best model for the monolingual scenario with 0.77, decreased noticeably for the code-mixed (achieving 0.63).

The work [172] also relies on sub-word embeddings. The architecture (SAEKCS), similar to [166], includes a CNN layer on top of the embeddings to extract local dependencies. Its output is processed by a BiLSTM layer after max-pooling, to learn long-term relations. On the top, a fully connected layer acts as the classifier. They evaluated SAEKCS using Kannada-English code-switching Youtube comments with an accuracy of 77.6%. They also assessed a sub-word LSTM (64.8%) and a BiLSTM (55.9%), suggesting that the short-term dependencies encoded by the CNN greatly benefit the model.

Summary Finally, we also have reviewed code-switching approaches for SMA. The more trending proposal is to use sub-word embeddings to allow guessing the meaning of unknown/out-of-vocabulary words. We have contrasted very different mixing languages, in the majority of cases, English mix. The next section is an overview of the deep learning implementations across the different setups.

2.4.4 An overview of the different deep learning implementations

Up to this point, we have reviewed a comprehensive corpus of research works that had leveraged deep learning for multilingual sentiment analysis. We had focused on the underlying hypothesis of the proposed approaches, highlighting what is common or different between the different solutions. Also, the results on the evaluation corpus.

In this section, we aim to make it easier for the reader to have a quick overview of the material we analyzed.

Table 2.1 lists the items of researches examined in this chapter, summarizing the architectures they used in their best model (Proposed Model), baselines in their experiments, distinctive ideas (Proposed Approach), the classification categories, multilingualism level, languages, and information about the corpus.

The classification categories mainly will be divided into positive or negative (binary), also with a neutral class (ternary) or five classes[6].

[6]Sometimes authors remove the *neutral* class or add another, e.g., *ambivalent*.

Table 2.1: Multilingual Approaches Used in DSA.

Proposed Model*	**Baseline Model***	**Proposed Approach**	**C[1]**	**L[2]**	**Language**	**Corpus**	**Reference**
CNN	translate-single-lang-CNN, multi-lang-CNN, multi-lang-without-identification-CNN, Random Forest	Trained large amounts of data in various languages (is trained for every single language), in three phases: unsupervised, distant supervised and supervised with multi-layer CNN	T	D & M	English, Italian, French, & German	Unlabeled tweets (300M), weakly-supervised data (40-60M) and annotated tweets (71K)	Deriu et al. (2017) [140]
CNN	**LSTM-Emb**, Conv-Emb, Conv-Emb-Freeze, Conv-Char & SVM	Cost-effective Character-based embedding and optimized convolutions	B	D	English, German, Portuguese & Spanish	Annotated tweets (128K, subset of 1.6M)	Becker et al. (2017) [141]
CNN	LSTM-Emb, Conv-Emb, Conv-Emb-Freeze, Conv-Char & SVM	Character-level embeddings with few learnable parameters	B	D & M	English, German, Portuguese & Spanish	Annotated tweets (128K, subset of 1.6M)	Wehrmann et al. (2017) [128]
CNN	word-Translation-CNN, char-Translation-CNN, 1-gram-SVM, 2-gram-SVM	Transformed characters into numbers corresponding UTF-8 decimal codes	B	D	English, Polish, German, Slovak, Slovenian & Swedish	Annotated tweets (150K, subset of 1.6M) & Labeled tweets (over 40 languages, 27.9K)	Zhang et al. (2017) [144]
CNN	**SVM**	N-gram bilingual mixed (English-French) input text source (based on a Naïve approach)	B	D & M	French, English & Greek	Labeled restaurant balanced reviews (62.6 K)	Medrouk & Pappa (2017) [147]
CNN	word-CNN, char-CNN, unicode-char-CNN, CuDNNLSTM (word and char)	Word-level & Character-level embeddings with two convolutional channels (one channel for each)	B	D	English, German, Portuguese, Spanish, Polish, Slovak, Slovenian & Swedish	Annotated tweets (193K, subset of 1.6 million)	Zhang et al. (2017) [146]

continued ...

Table 2.1: Multilingual Approaches Used in DSA (continued ...).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
BiLSTM	Average Skip-gram Vectors with LR & CNN-Subword-char-LSTM	Shared parameters of twin (siamese) networks with contrastive learning	T	M	English, Hindi	English annotated tweets (114K) & Hindi-English labeled sentences of Facebook posts (3.8K)	Choudhary et al. (2018) [163]
CNN	random initialized CNN, CNN + GloVe/FastText/Polyglot embeddings, regular CNN concatenate standard GloVe/FastText + multilingual sentiment embeddings (VADER, SocialSent or Amazon reviews), dual-channel-CNN + GloVe/FastText incorporates random initialized, Polyglot, VADER, SocialSent or static Amazon reviews embedding	Cross-lingual graph-based propagation (transfer-learning) from a rich source language with embeddings of supervised training on Amazon reviews to a dual-channel neural architecture	B	D	English, Spanish, Dutch, German, Russian, Italian, Czech, Japanese & French	Annotated movie reviews (12.2K, Rotten Tomatoes and AlloCine), labeled reviews (20.8K, TripAdvisor and Amazon Fine Food) & labeled tweets (4.8K)	Dong & De Melo (2018) [151]
CNN	LSTM (one-layer and two-layer), BiLSTM (one-layer and two-layer), CNN-LSTM, NB & SVM	Dictionaries of character and word indexes to produce code-mixed character and word embedding for a single NN	T	M	Bambara & French (mixed)	Labeled Facebook comments (17K, subset of 74K)	Konate & Du (2018) [164]
MNB + LSTM	Subword-LSTM	Ensemble of Multinomial Naïve Bayes with 1 and 2-gram features and many-to-one stacked LSTM over 3-gram encoding of sentences	T	M	English, Hindi	Hindi-English labeled sentences from Facebook posts (3.8K)	Jhanwar & Das (2018) [165]

continued ...

Table 2.1: Multilingual Approaches Used in DSA (continued ...).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
LSTM	**CNN**	N-gram raw corpus-based input, without any preprocessing, translation, annotation nor additional knowledge features	B	D & M	French, English & Greek	Labeled restaurant and hotel balanced reviews (91.8K)	Medrouk & Pappa (2018) [148]
BiLSTM	Doc2Vec + SVM, FastText & CNN	Embedding with an only distributed representation of the text	T	M	English, Bengali, Hindi & Kannada (English mixed)	Labeled sentences from Facebook comments (22.5K)	Shalini et al. (2018) [167]
BiLSTM	**SVM**	Learning new word embeddings based on limited training datasets and a pre-trained DNN exploiting transfer-learning from a rich source language with labeled data	B	D	English & Greek	Labeled TripAdvisor reviews (40K) & annotated tweets (480)	Stavridis et al. (2018) [152]
CNN + GAN + Attention-mechanism	SVM + Word2Vec, LSTM, CNN, NSC + UPA, UPNN	Combined CNN, GAN and user attention to learn specific and independent-language features from data with authorship information	B	D	English & Chinese	Annotated tweets (48.1K) & Weibo posts (53.6K)	Wang et al. (2018) [153]
GAN + DAN	DAN, mSDA, Machine Translation + DAN, CLD-KCNN, CLDFA-KCNN	Combined Bilingual Embedding (BWE), Deep Averaging Network (DAN) and adversarial training to learn independent-language features from a source language (English)	T, F	D	English, Chinese & Arab	English reviews from Yelp (700K), hotels reviews in Chinese (20K labeled / 150K unlabeled) & tweets in Arab (1.2K labeled)	Chen et al. (2018) [154]

continued ...

Table 2.1: Multilingual Approaches Used in DSA (continued ...).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
BiLSTM-CNN	CNN-hierarchical-BiLSTM, CNN-hierarchical-BiLSTM-gate-mechanism, LSTM, CNN, LSTM-attention-mechanism, CNN-attention-mechanism, RCNN-LSTM, hierarchical-LSTM, LSTM-sentences-relations	Word vector representation improvement based on gate mechanism, which obtains time-series relationship of different sentences in the comments through an RCNN, and gets the local features of the specific aspects in the sentence and the long-distance dependence in the whole comment through a hierarchical attention BiLSTM	B & T	D	English, Arabic, French & Chinese	Binary (400) and ternary (3.7K) labeled web reviews (4.1K)	Liu et al. (2019) [149]
BiLSTM-CNN	1-grams + 2-grams-SVM, 1-grams + 2-grams-NB-SVM, 1-grams + 2-grams-MNB, Tf-Idf-MNB, Lexicon Lookup, Char-LSTM, Subword-LSTM, FastText & SACMT	Hybrid architecture with sub-word level representations for the sentences, two parallel BiLSTM as Dual Encoder (Collective Encoder for overall sentiment and Specific Encoder with attention mechanism for sub-words) and linguistic features network	T	M	English, Hindi	Hindi-English labeled sentences of Facebook posts (3.8K)	Lal et al. (2019) [169]
BiLSTM	SVM-MONO, SVM-MT, ARTEXTE-SVM-based, ARTEXTE-Ensemble, BARISTA-SVM-based, BARISTA-Ensemble, BLSE, BLSE-Ensemble & BiLSTM-MT	Low-resource language embeddings + mapping function joined with rich-resource language embedding through k-NN refinement, BiLSTM as encoder layer, then fully-connected layer with softmax for prediction	F-1	D	English, Spanish & Catalan	English labeled tweets (33.7K), Spanish OpeNER (1.3K) & Catalan MultiBooked (1K)	Jabreel et al. (2019) [159]

continued ...

Table 2.1: Multilingual Approaches Used in DSA (continued ...).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
Double LSTM	Double LSTM with several combinations of optimizers and loss functions & Subword-LSTM	Low-resource + code-mixed corpus to train embeddings. A joint feature of sentences (Subword + word levels), preceded by Double LSTM layer	T	M	English & Hindi (mixed)	Hindi-English labeled sentences of Facebook posts (3.8K)	Mukherjee (2019) [170]
LSTM-AAE-BiGRU	MT-SVM, MT-BiGRU & TL-BiGRU	Contextual word embeddings (Word2Vec+LSTM, source and target languages), AAE to train the transformation matrix and calculate the cross-lingual word embeddings (average), then, transfer the sentiment classifier (BiGRU)	B	D	English, Chinese & German	Amazon labeled documents (28.9K) and unlabeled documents (80K) for each pair of language category (books, DVD, music)	Shen et al. (2020) [157]
BiLSTM + Attention-mechanism	NB, SMO (SVM), RF, BiLSTM-CNN, Double BiLSTM, GloVe + BiLSTM-CNN, GloVe + Double BiLSTM, GloVe + Attention-based-BiLSTM, $\text{BERT}_{base-uncased}$ & $\text{BERT}_{base-domain-uncased}$	Attention mechanism to extract such words that are important to the meaning of the sentence and aggregate the representation of those informative words to form the sentence vector; a sigmoid layer is used to predict the correct label	T	M	English, Hindi & Bengali (English mixed)	English, Bengali-English, Hindi-English annotated tweets (9.2K, 5.5K, 18.4K) & Hindi-English labeled sentences of Facebook posts (3.8K)	Jamatia et al. (2020) [171]
BiLSTM	LASER-CNN, FastText-BiLSTM & fastText-CNN	Transfer learning by LASER with (low-resource) language corpus, BiLSTM, then predict the sentiment of texts in other (high-resource) language	F + 1	D	Polish, Dutch, English, French, German, Italian, Portuguese, Russian & Spanish	Online medicine, hotels, school, products reviews (8.4K for each language)	Kanclerz et al. (2020) [161]

continued ...

Table 2.1: Multilingual Approaches Used in DSA (continued ...).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
CNN-BiLSTM	NB, BiLSTM & Subword-LSTM	Three stages classification with subword embeddings + CNN-BILSTM: first positive or not, then negative or not, and then, computed classification matrix of them	T	M	English & Kannada (mixed)	Annotated Youtube comments (10.4K)	Chundi et al. (2020) [172]
LSTM-CNN	BilBOWA + CNN, VecMap + CNN, BilBOWA + LSTM, VecMap + LSTM, BilBOWA + CNN-LSTM, VecMap + CNN-LSTM & BilBOWA + LSTM-CNN	Trained bilingual embeddings (VecMap, on one high-resource and another low-resource language), used to on target language (low-resource), followed by a DL classifier for predict polarity	B	D	English & Persian	Binary (11K, Persian Digikala reviews), five categories (200K, English Amazon reviews)	Ghasemi et al. (2020) [162]

* **Bold**: model with best performance.
[1] Classification (**C**): **B** (Binary), **T** (Ternary), **F** (Five categories).
[2] Multilingualism Level (**L**): Document (**D**), Mix (**M**).

2.5 Discussion

In this section, we discuss the main findings of our study. We also highlight some unexplored topics that may hint at interesting directions for further research.

2.5.1 Languages and social media in MSA

As showed in Table 2.1, proposed models have applied different DLs architectures for tackle MSA, with good results (see section 2.4). First of all, we cannot compare the results over this task, because there are numerous models that have been proposed on various corpora across different social media platforms. In essence, DL models aided to reduce the load of translation that is usually required to perform MSA, which is one of the aims of language-agnostic MSA. There are not many available corpora for this task, mainly for SA on low-resource languages. However, some works tried to solve this issue with their proposals [140, 159], following the zero-shot learning principle for a multilingual classifier, relied on high-resource languages datasets.

The 24 studies we analyzed covered 23 different languages. In most cases English was the resource-rich language, except [161] and [164]. However, authors have explored synergies between different languages, as shown in Figure 2.1. Concerning the social media Twitter and Miscellaneous[7] account for most of the works. Figure 2.2 the data de-aggregated by language and social media.

Other dimensions for interesting analysis are the relation between the MSA setup and the architecture proposed to deal with the problem. The next subsection is devoted to this point.

2.5.2 DL architectures for MSA

A comparison between the backbone of the different architectures suggests that in general, for multilanguage sentence-level SA, authors have explored a plainer architecture. It leverages trainable embeddings preceding the feature extractor and finally a classifier layer. Regardless, there is a wide range of alternatives in the design of the classifier, from a single BiLSTM [152] to parallel CNN [146]. Only one study focused on the aspect-level SA. Based on the limited number of samples, it can be concluded that this less exploited area is a niche for future work. So, it may be difficult to draw conclusions related to the best architectural decisions to tackle this problem. However, experiments in [149] suggest that the attention mechanism as well the aspect embedding plays a key role. Moreover, results from another work [173] show that the mere addition of attention to a simpler model such as LSTM or CNN does not lead to state-of-the-art results. This is indicative of the convenience of combining feature extractors at different levels for aspect-based SA.

For cross-lingual SA network designs are more diverse. However, two core ideas can be identified. The use of an adversarial module to drive the feature extractors to language-independent representations as in [153] or [154]. The other is leveraging trainable cross-lingual embeddings or even pre-trained ones such as Facebook LASER [161].

Code-switching backbones tend to be simpler than cross-lingual, yet more elaborated than the multilingual ones. They can include parallel encoders at a character, word, or subword levels [169, 170] or implement ensemble models including deep neural networks and other classification techniques [165]. These and other findings are shown in Figure 2.3.

In general, a lot of effort has been devoted to comparing feature extractors based on CNN, LSTM, or BiLSTM and different levels of embeddings. We analyzed the co-occurrences of the different types of networks, attention mechanisms, etc. within the same

[7] Under this category, we considered works such as [149] which used the SemEval 2016 Task 5, and other corpora that aggregated text from different sources.

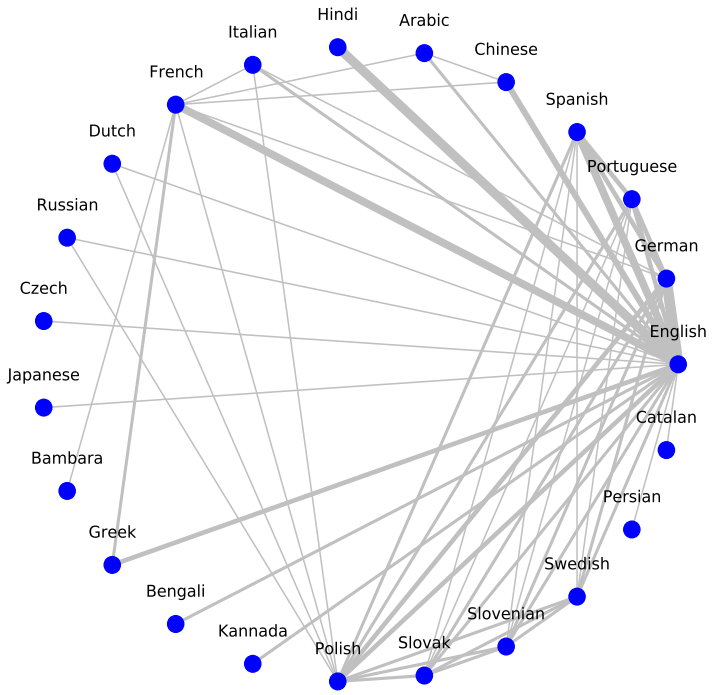


Figure 2.1: Synergies between languages. A link between two languages indicates that both has been used simultaneously in a model, as source or target language, or to learn multilingual feature spaces.

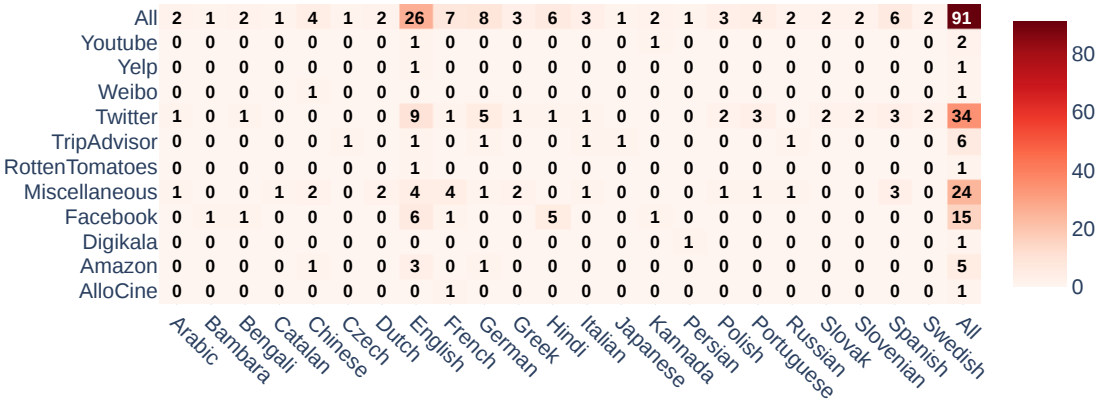


Figure 2.2: Languages vs. Social Media.

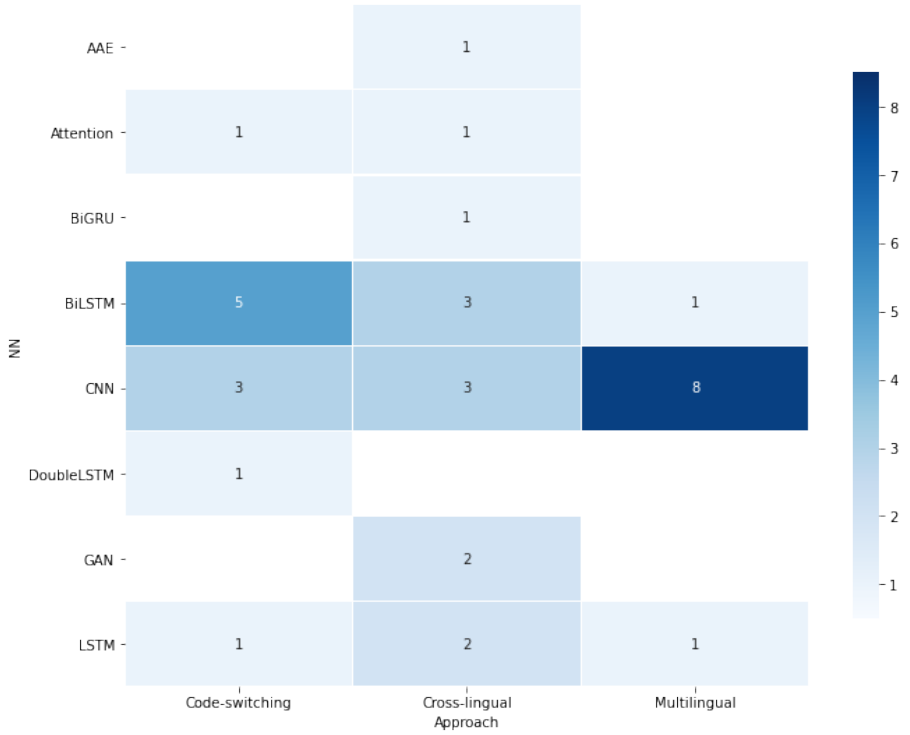


Figure 2.3: Neural Network (NN) architectures vs. approaches and topics.

model, to visualize how authors have been using them. Edges in Figure 2.4 means that the two concepts at nodes have been used together in a model, the weight, how often this relationship has occurred. The graph shows how CNN is preferred along with hybrid architectures, where CNN + BiLSTM is the most studied with [164] and [167] reporting comparisons between models where they only changed this setting, resulting in better performance for the CNN models. Instead, results were equivalent in [148] and better for LSTM in [141]. Thus, seem to be no consensus about this subject.

2.5.2.1 Embedding approaches for MSA

Most authors rely on some embedding types, with a trend towards learning the embedding along the training process. However, there is no common opinion about whether to work at a character, word, or document level. Authors such as [141] advocate character level embeddings due to their simplicity. However, they report slightly better results for the word level case but improved in [128], which shows the character level helpful for achieving a language-independence. However, in [164] they compared the character and word-based embeddings, with the latter yielding better results. In addition, sub-word embeddings reported outperforming the character ones in [169].

Not, surprisingly we can find pre-trained sub-word multilingual embeddings, such as [174], which be useful for MSA and SA for low-resource languages. However, there are another alternatives to be explored such as document level [167], a combination of different levels [146], [149] and even sentiment-driven embeddings [151], universal embeddings [159] or the use of some tools, such as LASER [161].

2.5.3 Suggestions for future research

Finally, we elaborate on the current state of research and provide a pathway for what can be done or needs to be done within the following few years.

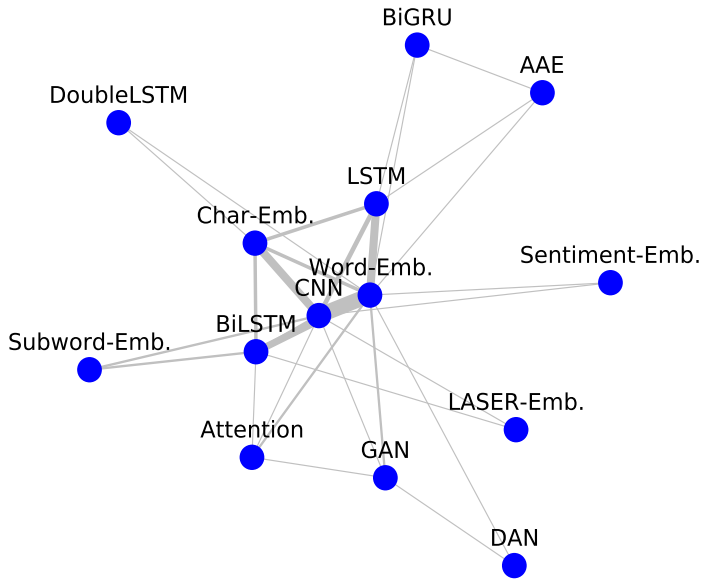


Figure 2.4: Neural Network architectures and its relations, out of 24 , across reviewed papers.

Little-explored MSA levels Hitherto, Aspect-Based Sentiment Analysis has not widely been addressed using multilingual deep learning approaches. As [149] suggests, tackling this problem may require more complex architectures. Moreover, it needs to be studied if current proposals can handle mixing setups such as aspect-based code-switching.

MSA setup shift across time Figure 2.5 suggest a shift of the interest from multilingual to cross-lingual and code-switching approaches. In MSA this can be explained since initially most of the works focused on the multilingual setup evaluating many variations of the same design. Researchers could perceive this path as depleted. Also, the adoption of transformer-based architectures such as BERT [12] and even multilingual models such as *Multilingual BERT*[8] [12] allows the researchers to focus on fine-tuning the models instead of training them from scratch with a multilingual corpus as has been typical for the multilingual setup.

Multilingual representations: Multilingual embeddings and adversarial training seem to be the most common approaches to achieve multilingualism within the analyzed corpus. But there is not a common standpoint about the level of embedding to use or how a single model can encode a multilingual or language-agnostic feature space useful for the downstream tasks. However, this debate seems to be shifted to the transformer-based architectures where different tokenizers are being considered [77]. Moreover, despite the success in training transformers in a multilingual corpus, recent studies suggest that there is a lot of room for improvement [175]. In this sense probably we will see an increased number of works studying the impact of the differences between languages and language families.

SA-specific representations For MSA, the aforementioned architectures would handle the specifics of this domain such as the code-switching or the aspect-based setup. It will

[8] `https://github.com/google-research/bert/blob/master/multilingual.md`

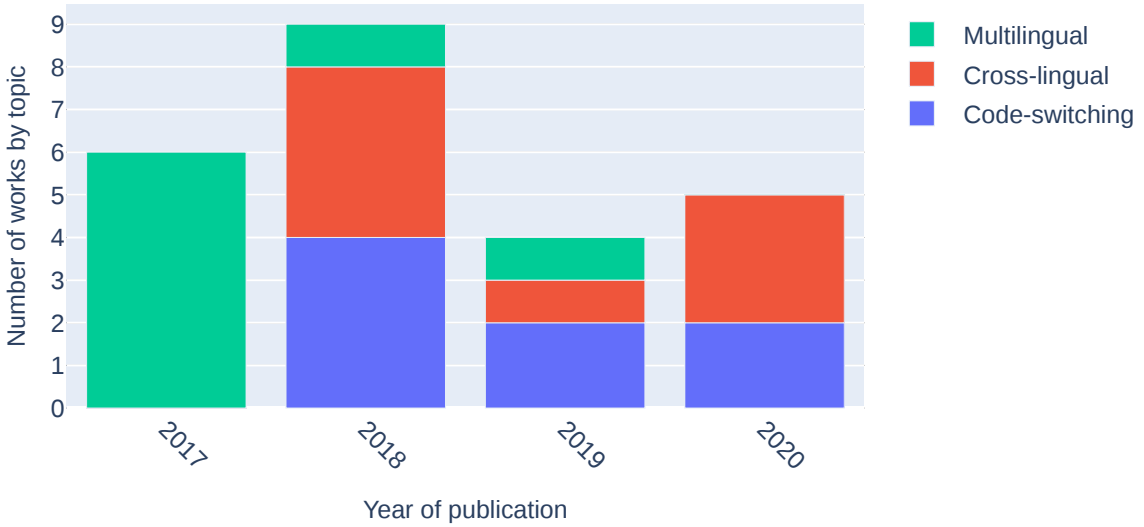


Figure 2.5: MSA approaches, out of 24 papers, across fours years.

be necessary to study if it is worth to couple techniques such as attention between different levels of representation, sentiment embeddings, or adversarial learning (e.g GAN-BERT [176]) into the aforementioned architectures.

Low-resource languages and dialects Despite languages from different families having been studied (see Figure 2.2) the coverage is far from complete. Moreover, the steady interest in sentiment analysis, the lack so far of a universal approach, and the new opportunities [177, 178] would trigger the development of systems and corpora for SA in other languages. In this sense, we would see tailored solutions dealing with dialects and mixing languages. Besides India (native languages mixing with English) there are other large groups such as (a) Mexico and USA (Spanglish), (b) Brazil and its border countries, Portugal and Spain (Portuñol) (c) Paraguay (Jopara[9], Portuñol), to mention few cases. Nevertheless, the scarce of available corpora is a challenge for tackling these code-mixing tasks. For instance, in [171] we could observe that with an English mixing MSA setup, BERT had been unable to outperform the traditional DL models. Thus, substantial progress needs still to be made.

Languages varieties Multivariate or multi-dialect sentiment analysis as MSA, aim to classify the polarity but on the dialects of certain language, often systems must be trained with one or more language variants and tested with a different language variant. Some examples of multi-dialect SA, are given in Algerian [179], Arabic [180], Spanish [181, 182] or Tunisian [183]. These works demonstrate the inception of multi-dialect SA as a variant of MSA that further much attention.

2.6 Conclusion

In this chapter, we reviewed 24 works that studied 23 different languages and 11 sources. The observed trend evidences the steady interest in this domain, so we expect to see this direction continue.

As regards the different MSA setups, the multilingual approach seems to be of decreased interest. However, aspect-based sentiment analysis is still an understudied domain and an open research field with a lot of scope for future works.

[9]Mixing Guarani (an indigenous language) with Spanish.

We highlighted the main ideas authors proposed to tackle the challenge that represents the lack of annotated data or to achieve language independent models. Despite state-of-the-art results in some cases, the simpler backbone comprising embeddings, a feature extractor, and a classifier seems to be unappropriated for more complex scenarios. Also, there are unsolved questions such as which type of embedding captures better the particulars of MSA. We hint about future research directions, for example, if ideas such as contextualized embeddings, which have proven very useful in other tasks, can further improve MSA. Finally, although studies have covered very different languages such as Arabic, Chinese, or Hindi, the world is extraordinarily rich in languages, cultures, and ways of expressing feelings. Thus, better approaches need to be assessed or developed for new scenarios.

Acknowledgements

We are immensely grateful to David Vilares (Universidade da Coruña, Spain), for his very valuable review of this chapter.

Chapter 3

Data Collection: Creating Corpora for Low-resourced Languages and Code-switching [1]

The chapter is devoted to the creation of corpora for low-resourced and code-switched languages. It covers the (i) collection of COVID-19 related tweets in Spanish and (ii) the text data collection effort for sentiment analysis of Guarani-dominated tweets.

3.1 Introduction

The datasets created in this chapter coming from 'real data' and from two very different contexts, a priori. It should be noted that all these data sources have their own particularities, specific to their context, they present different forms of language, communication, vocabulary, etc., such as (a) one of a more of a social impact, the COVID-19 (topics discovering on Twitter.com), where there is usually no entity with an economic interest, but rather it is the Spanish society itself that is interested: it wants to be alert, prevented against possible outbreaks of this disease that could be catastrophic for many people, where the beneficiaries are the citizens as a whole; and (b) a second context, opinions written in truly low-resource and code-switching languages, where getting enough texts to perform an analysis and modeling is a challenge. In this respect, we provide resources that can be used directly by both academia and industry interested in analyzing texts in Guarani (a South American indigenous language spoken mainly in Paraguay) and/or Jopara (i.e., the mixture of Guarani and Spanish languages), being our proposal the first Guarani-dominant one.

Thus, we summarize in this chapter our effort to compile three different datasets for topic and sentiment analysis in social media. All of them are created from posts written in low-resource languages, namely Spanish (the variety spoken in Spain), Guarani and Jopara. Note that our research objective is focused on the application of machine learning to this kind of languages and as already seen in the last chapter (§2), it is not easy to find datasets for them.

We first collect around five million tweets about COVID-19, mostly between 1 January of 2019 to 20 April of 2020, focusing on the possibilities of performing effective and representative topic modeling over this particular large set of Spanish tweets. It is worth noting that most of the efforts focused on specific languages (mainly rich-resource ones) [184, 185, 186],

[1] The main content of the chapter has been extracted from the papers published in *Procesamiento Del Lenguaje Natural* [95] (see section A.3), *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* (co-located in *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*) [96] (see section A.5) and another paper to be submitted to a journal.

but not on language varieties, as is the case of our proposed dataset. More particularly, tweets written in the variety of Spanish spoken in Spain. This task, which involves the discovery of topics in Spanish, led us to create an unlabelled dataset to perform it. We know of few works in this context [187, 188], but as our aim was to study opinions mainly at the onset of the COVID-19 pandemic, such available datasets were not suitable for our task.

Second, we collect a corpus for polarity classification of Jopara tweets, which mixes Guarani and Spanish languages, being the former the dominating language in the corpus. We also discuss the difficulties that we had to face when creating such a resource, such as finding enough Twitter data that shows sentiment and contains a significant amount of Guarani terms (contraries to [40], which were construct a Spanish dominant Jopara corpus).

As mentioned above, we are not aware of datasets that are suitable for conducting sentiment analysis on Guarani-dominated texts. Therefore, we also create other corpora, extending the task of sentiment analysis along some dimensions. We present a multi-dimensionally annotated affective corpus containing Jopara code-switching tweets, where Guarani is the dominant language. More specifically, we consider the following dimensions: emotion recognition, humor detection and, identification of offensive language.

Contribution Our contribution are two fold. First, we collected a Spanish geolocated corpus to perform unsupervised learning tasks, like that the topic model, in an current dataset, due to the current pandemic we are suffer. Second, we create several corpora to perform sentiment analysis and affective computing in Guarani/Jopara, in a truly low-resource scenario, where the dominant Guarani corpora created are the first of this kind.

3.2 Methods

In this section, we describe the creation process of our Twitter data collection, which aims to create corpora for low-resourced languages and code-switching setups.

3.2.1 COTwes: COVID-19 outbreak's Tweets related written in Spanish - Spain variety

In what follows, we describe our methodology, decomposed into two steps: (i) the collection of the corpus, (ii) the language identification and the geolocation of the tweets. It should be noted that our objective was to collect tweets related to the COVID-19 outbreak, limiting them to tweets written in the varieties of Spanish spoken in Spain.

Collection of tweets We first defined a set of keywords to download relevant tweets: *coronavirus*, *COVID-19*, *COVID19*, *2019-nCoV*, *2019nCoV*. Further, as of March 3th, 2020 we added more keys: *SARS-CoV-2*, *SARSCoV2*, *CoV-19*, *CoV19*, *COVD19*, *COVD 19*, *corona virus*, *corona outbreak*.

More particularly, we collected a multilingual corpus of 32.68M tweets, including Twitter posts from 1 January of 2019[2] to 20 April of 2020; from all over the world.

We scraped the tweets using the `GetOldTweets-python3` (GOT3) library.[3] The reason to use this tool was that it allowed to retrieve old tweets without time limitation. However, the tool did not permit us to filter the retrieval by language. Besides, the Twitter Official API cannot retrieve tweets more than a week ago with a free subscription mode.[4]

[2] In order to have some preceding context, but expecting just to be able to retrieve a small number of tweets.

[3] `https://github.com/Jefferson-Henrique/GetOldTweets-python`

[4] However, we noted that GOT3 as of 18 September 2020 has been suspended due to the new Twitter

Language identification and geolocation The next step is language identification to keep only the Spanish tweets. We used four tools for detecting languages, since with GOT3 we could not obtain the language attribute. Those four tools were: `polyglot`,[5] `langdetect`,[6] `langid.py`,[7] and `fastText`[8] [189]. The language is assigned based on majority voting. In case of a tie, we consider the tweet to be Spanish, except if all tools predicted a different language.

In total, we identified 5.35M Spanish tweets. In this work, we try to restrict the analysis to the content generated in Spain. For this purpose, we proceeded to filter the tweets in Spanish using the location attribute of the user profile, and look for the name of Spanish cities with more than 50K inhabitants, province names, autonomous regions names, and also any location specified as simply 'Spain'.[9]

Dataset and limitations After the cleaning process, we obtained ∼1.85M tweets (∼1M deduplicated tweets by text content) for our topic modeling analysis. It is fair to point out that there is a percentage of tweets with a risk of not being correctly filtered, since the same place name might exist in more than one Spanish speaking country (e.g., 'Guadalajara' for Spain vs. 'Guadalajara' for Mexico). This is a common limitation on Twitter analyses, when it comes to analyze geolocated tweets (see for instance [190]).

3.2.2 JOSA: The Jopara Sentiment Analysis dataset

In what follows, we describe our attempts to collect Jopara tweets. Note that ideally we are interested in tweets that are as Guarani as possible. However, Guarani is intertwined with Spanish, and thus we have focused on Jopara, aiming for Guarani-dominant tweets, in contrast to [40]. We found interesting to report failed attempts to collect such data, since the proposed methods would most likely work to collect data in rich resource languages. We hope this can be helpful for other researchers interested in developing datasets for low-resource languages in web environments.

In this line, Twitter does not allow to automatically crawl Guarani tweets, since it is not included in its language identification tool. To overcome this, we considered two alternatives: (i) using a set of Guarani keywords (§3.2.2.1), and (ii) scrapping Twitter accounts that mostly tweet in Guarani (§3.2.2.2).

3.2.2.1 Downloading tweets using Guarani keywords - An unsuccessful attempt.

As the Twitter real-time streamer can deal with a limited number of keywords, we consider 50 different keywords which are renewed every 3 hours, and used them to sample tweets. To select such keywords, we considered two options:

1. *Dictionary-based keywords*: We used 5.1K Guarani terms from a Spanish-Guarani word-level translator.[10] We then downloaded 2.1M tweets and performed language identification with three tools: (i) `polyglot`,[11] (ii) `fastText` [189] and (iii) `textcat`.[12]

policies on tweet payload.

[5] `https://polyglot.readthedocs.io/en/latest/Detection.html`

[6] `https://pypi.org/project/langdetect/`

[7] `https://pypi.org/project/langid/`

[8] `https://fasttext.cc/docs/en/language-identification.html`. We used the large model.

[9] We obtained the list of place names from the Instituto Nacional de Estadística (INE), `https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525`.

[10] `https://github.com/SENATICS/traductor-espanhol-guarani`

[11] `https://polyglot.readthedocs.io/en/latest/Detection.html`

[12] `https://www.nltk.org/_modules/nltk/classify/textcat.html`

We assume that the text was Guarani if at least one of them classified the text as Guarani. After this, we got 5.3K tweets. Next, a human annotator was in charge of classifying such a subset, obtaining that only 150 tweets, over the initial set of 2.1M samples, were prone to be Guarani-dominant.

2. *Corpus-based keywords*: We first merged two Guarani datasets[13] [191], that were generated from web sources and included biblical passages, wiki entries, blog posts or tweets, among other sources. From there, we selected 550 terms, including word unigrams and bi-grams with 100 occurrences or more. Again, we downloaded tweets using the keywords and collected 7M of tweets, but after repeating the language identification phase of step 1, we obtained a marginal amount of tweets that were Guarani-dominant.

Limitations This approach suffered from a low recall when it came to collecting Guarani-dominant tweets, while similar approaches have worked when collecting data for rich-resource languages, where a few keywords were enough to successfully download tweets in the target language [32]. In this context, even if tweets contained a few Guarani terms, there were other issues: (i) words that have the same form in Spanish and Guarani such as '*mano*' (*'hand'* and *'to die'*), (ii) loanwords,[14] such as '*pororó*' ('*popcorn*') or *'chipa'* (traditional Paraguayan food, non-translatable); (iii) or simply tweets where the majority of the content was written in Spanish. Overall, this has been a problem experienced in other low-resource setups [193, 194], so we decided instead to look for alternatives to find Guarani-dominant tweets.

3.2.2.2 Downloading tweets from Guarani accounts - A successful attempt.

In this case, we crawled Twitter accounts that usually tweet in Guarani.[15] We scrapped them, and obtained more than 23K Guarani and Jopara tweets from a few popular users (see Appendix C). Using the same Guarani language identification approach as in 1, we obtained 8,716 tweets. To eliminate very similar tweets that could contaminate the dataset, we removed tweets with a similarity greater than 60%, according to the *Levenshtein* distance. After applying this second cleaning step, we obtained a total of 3,948 tweets.

Annotation The dataset was then annotated by two native speakers of Guarani and Spanish. They were asked to: (i) determine whether the tweet was strictly written in Guarani, Jopara or other languages (i.e., if the tweet did not have any words in Guarani); and determine whether the tweet was positive, neutral or negative. For sentiment annotations consolidation, we proceeded similarly to the SemEval-2017 Task 4 guidelines [16, § 3.3].[16] See subsection B.1.1 for more details about the process. We then filtered the corpus by language, including only those labeled as Guarani or Jopara, to ensure the samples are Guarani-dominant. This resulted into 3,491 tweets.

Limitations Although this second approach is successful when it comes to collecting a reasonable amount of Guarani-dominant tweets, it also suffers from a few limitations. For instance, the first part of Table 3.1 shows that due to the nature of the crawled Twitter accounts (who tweet about events, news, announcements, greetings, ephemeris, tweets to encourage the use of Guarani, etc.), there is a tendency to neutral tweets. Also, as the

[13] BCP-47 *gn* and *gug* codes.

[14] Frequent in Paraguay and border countries [192].

[15] We followed `http://indigenoustweets.com/gn/`. We did not use an external human annotator as in 1, since the crawled accounts tend to tweet in Guarani.

[16] We obtained a slight agreement following Cohen's kappa metric [195].

number of selected accounts was small, the number of discussed topics might be limited too. We comment on this a bit further in the section C.

Balanced and unbalanced versions As we are interested in identifying sentiment in Jopara tweets, we also created a balanced version of JOSA. Note that unbalanced settings are also interesting and might reflect real-world setups. More particularly, we split each corpus into training (50%), development (10%), and test (40%). We show the statistics in Table 3.1.

For completeness, in Table 3.2 we show for the balanced corpus the top five most frequent terms (we only consider content tokens) for Guarani, Spanish and some language-independent tokens, such emoticons. This was done based on manual annotation of a Guarani-Spanish native speaker.

Version	**Total**	**Positive**	**Neutral**	**Negative**
Unbalanced	3,491	349	2,728	414
Balanced	1,526		763	

Version		**Train**	**Development**	**Test**
Unbalanced	3,491	1,745	349	1,397
Balanced	1,526	763	152	611

Table 3.1: JOSA detailed statistics and splits for the un/balanced versions.

Category	**#Terms**	**Most frequent**
Guarani	4,336	guaranime, ñe'ẽ, mba'e, guarani, avei
Spanish	1,738	paraguay, guaraní, no, es, día
Other*	1,440	alcaraz, su, rt, juan, francisco
Mixing	368	guaraníme, departamento-pe, castellano-pe, castellanope, twitter-pe
Emojis	112	🇵🇾 😂 👏 xD :)

*We include reserved words, proper nouns, acronyms, etc.

Table 3.2: Frequent terms for the balanced JOSA.

3.2.3 JOTAD: The Jopara Text-based Affect Detection dataset

We now describe the steps that we followed to create our collection for Jopara text-based affect detection. Two are the challenges that we need to face in the context of an endangered language such as Jopara/Guaraní: (i) finding tweets that are not highly contaminated by Spanish, with which Guaraní is highly intertwined, and (ii) finding enough tweets that are relevant for affective tasks such as the ones addressed in this work.

Collection Twitter does not include Guarani in its language identification tool, and therefore we cannot directly monitor a stream of tweets published in this language. To offset this problem, we employed a set of Guarani keywords from JOSA dataset (see subsection 3.2.2), in particular, words from the positive and negative polarity classes. That is, we tokenized the tweets of both classes and represented them as a 1-hot vector of uni-grams with a Term-Frequency scheme.

We then selected a total of 763 JOSA's Guarani terms as keywords, including only uni-grams of words with 3 or more occurrences. We considered the 50 different keywords that are rotated every 3 hours, and used them to sample tweets. Note that the real-time Twitter streamer can only deal with a limited number of keywords. Tweets were downloaded for 6 months.

Downloading We downloaded 7.4M tweets and performed language identification with the same three tools used in subsection 3.2.2, i.e., (i) `polyglot`,[17] (ii) `fastText` [189] and (iii) `textcat`.[18] We assumed that the text is Guarani if at least one of them classified it as such. Following this, to exclude very similar tweets that could contaminate the dataset, we removed tweets with a similarity of more than 70%, according to the *Levenshtein* distance. After this second cleaning step, we obtained 2.8K Guarani-dominant tweets.

Limitations This approach suffered from low recall when it came to collecting Guarani-dominant tweets, which has already been a problem experienced with Guarani (subsection 3.2.2,[196]). However, we decided to take this reasonable number of Guarani-dominant tweets, due to the low-resource nature of Guarani, which also suffers from high code-switching with Spanish.

Annotation The dataset was annotated by two native speakers of Guarani and Spanish, one female and one male, both aged 27-35. They were asked to:

1. classify the predominant emotion of the tweet, i.e., happy, sad, angry and other (tweet showing none of the aforementioned emotions or no emotion - neutral emotion).

2. classify the tweet as toxic or offensive, i.e., hate speech, profane or inappropriate language, insults, or threats); and

3. determine whether or not the tweet was funny, i.e., if the content was intended to be humorous.

As in JOSA dataset (subsection 3.2.2, they also were asked to determine whether the tweet was strictly written in Guarani, Jopara or other languages (i.e., if the tweet did not have any Guarani words). See subsection B.1.2 for more details about the process. We then filtered the corpus by language, including only those labeled as Guarani or Jopara, to ensure the samples will be Guarani-dominant. This resulted in $2,364$ tweets.

Guidelines and inter-annotator agreement For emotion recognition annotations, we proceeded similarly to the SemEval-2019 Task 3 guidelines [17]. We obtained a moderate agreement (0.55) following Cohen's kappa metric [195, Figure 1;p. 576].

In the same way, for humor detection we followed [197, 30] shared-tasks[19] and also obtained a moderate agreement (0.52).

In the case of offensive language identification, we followed SemEval-2020 Task 12 [32] work and obtained a more good Cohen's kappa metric than emotion recognition and humor detection, we got a substantial agreement (0.73).

Dataset We then created three new corpora for Jopara affect detection at sentence level:

1. the Jopara emotion recognition (JOEMO),

2. the Jopara humor detection (JOFUN) and,

3. the Jopara toxic and ofensive language identification (JOFF+).

[17] `https://polyglot.readthedocs.io/en/latest/Detection.html`
[18] `https://www.nltk.org/_modules/nltk/classify/textcat.html`
[19] HAHA 2018 and 2019 competition editions and SemEval-2020 Task 7, respectively.

We split each corpus into training (64%), development (16%), and test (20%). Table 3.3 shows the number and amount of tweets by classes corresponding to each subset and corpus. It is important to highlight that the dataset is unbalanced, where the number of tweets per category does not follow the same distribution.

The Table 3.3 shows additionally the number of tweets from each class and from each corpus, which was annotated as Guarani or Jopara. This is based on the annotation process referred in section 3.2.3.

Class	gn	jopara	Total	Train	Development	Test
			JOEMO			
`e_angry`	166	315	481	315	75	91
`e_happy`	163	194	357	216	52	89
`e_other`	305	336	641	411	111	119
`e_sad`	50	42	92	63	13	16
Total	684	887	**1,571**	1,005	251	315
			JOFUN			
`fun`	145	359	504	323	75	106
`no_fun`	668	670	1,338	855	220	263
Total	813	1,029	**1,842**	1,178	295	369
			JOFF+			
`off`	110	239	349	232	47	70
`no_off`	833	988	1,821	1,156	301	364
Total	943	1,227	**2,170**	1,388	348	434

Table 3.3: JOTAD statistics and splits for the JOEMO, JOFUN and JOFF+ corpora.

Importantly, we have created a multi-annotated sentiment analysis dataset, a task that is not very common for low-resource languages. Figure 3.1 shows the overlap of tweets in the different corpora.

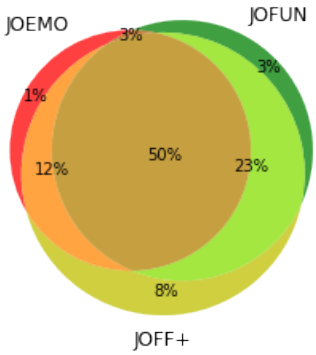


Figure 3.1: Venn diagram showing the overlap between our three corpus.

3.3 Results

The workflow presented for the creation of datasets for low-level resource languages provides a material basis and a test bed for the construction of new NLP resources such as these. We have covered findings, difficulties and successful as well as unsuccessful attempts. On the other hand, we aim to contribute substantially to corpus development, especially in the creation of corpora for low-resource languages. This chapter discusses fundamental issues and important aspects such as capturing tweets and their effective downloading, corpus size, representativeness, balancing, sampling filtering and even tagging and annotation, as well as peripheral issues such as mixed-coded corpora.

Table 3.4 shows the statistics of the Twitter data collection effort results. The tweet IDs[20] are available at the links listed in Table 3.4.

Corpus	Total	Classes	Task	Language
COTwes[1]	$1.85M$	-	Unsupervised learning	Spanish
JOSA - un/balanced version[2]	$3,491$ $1,526$	positive, neutral, negative	Polarity classification	
JOEMO[3]	$1,571$	angry, happy, sad, other	Emotion recognition	Guarani/Jopara
JOFUN[3]	$1,842$	fun, not fun	Humor detection	
JOFF+[3]	$2,170$	offensive, not offensive	Offensive language identification	

[1] https://doi.org/10.7910/DVN/6PPSAZ

[2] https://doi.org/10.7910/DVN/GLDX14

[3] https://github.com/mmaguero

Table 3.4: Twitter data collection statistics.

3.4 Conclusion

In subsection 3.2.1 we collected a large number of tweets about COVID-19 using keywords and cleaned them to keep only Spanish tweets that were written in Spain.

On the other hand, in subsection 3.2.2, we collected the first Guarani-dominant dataset for sentiment analysis (polarity classification), and described some of the challenges that we had to face to create a collection where there is a significant number of Guarani terms. Lastly, we collected the first dataset for affect detection in a multiple annotation task. More particularly, we covered the emotion recognition, the humor detection and the offensive language identification. It should worth that this corpora are so important to perform a less biased affective computing on social media in a truly low-resource language, even if it suffer of code-mixing.

The corpora created play an essential role in natural language processing (NLP) research as well as a wide range of socio-linguistic investigations. In the next chapters, we will describe our approaches and study for topic and sentiment analysis with the presented corpora in this chapter.

Acknowledgements

We thank the annotators that labelled JOSA as well as JOTAD. We also thank ExplosionAI for giving us access to the Prodigy annotation tool[21] with the Research License.

[20] Contact the author to obtain a version with text content for research purposes.

[21] https://prodi.gy/

Chapter 4

Topic Modeling Approach for Spanish: What Twitter Users Discussed About the COVID-19 Outbreak in Spain [1]

In this chapter, we automatically extracted topics to capture what Twitter users in Spain were discussing during the beginning of the COVID-19 pandemic (with the collected data in the subsection 3.2.1). We also reviewed related works. Here we introduced a discriminative route, extracting the most prominent words and phrases. We perform an in-depth qualitative analysis of these topics, as well as a small quantitative evaluation framework based on a human evaluation.

4.1 Introduction

The outbreak of the SARS-CoV-2 virus and the global spread of the COVID-19 disease has encouraged people and organizations to express their opinion, discuss topics and warn about the evolution of the pandemic on social media platforms such as Twitter.

Unlike previous occasions, such as SARS-CoV in 2002 or MERS-CoV in 2012 [198], where social media still were in an embryonic state and natural language processing (NLP) still had limited practical applications; we are now in a situation where users generate a vast amount of written content, that can be analyzed by automatic tools to discover the topics societies care about, and their sentiment. This has been already the case for some precedent events or catastrophes in recent years, such as the 2016 US political elections [199] or some natural disasters, such as the 2011 East Japan Earthquake [200].

In relation to the COVID-19 pandemic, a few specific NLP workshops [201, 202] have already attempted to highlight how NLP can be used to respond to situations like the current one; addressing a number of challenges that include mining scientific literature and social media analysis, among many others [203, 204, 205]. With research purposes, there has been also efforts on releasing NLP datasets discussing COVID-19 topics [206, 207, 208]. In this context, the area of topic modeling has not been a stranger to this problem, and a number of authors have shown the options that clustering online posts such as tweets or Facebook messages can offer to monitor and evaluate the evolution of the pandemic through time [184, 185, 186].

[1] Chapter based mainly on the content of a paper published in *Procesamiento Del Lenguaje Natural* [95] (see section A.3).

Contribution In this chapter, we also focus on the possibilities of performing effective and representative topic modeling over a large set of Spanish tweets (collected in subsection 3.2.1), which consist of around five million tweets about COVID-19, mostly between 1 January of 2019 to 20 April of 2020. Then, we apply latent Dirichlet allocation (LDA) [51] to compute relevant topics in an unsupervised way and obtain meaningful keywords and sentences through generative and discriminative routes. Finally, we provide an analysis to shed some light on the quality of the extracted topics, and how faithfully they represent what was happening in the Spanish society at different moments of the pandemic.

4.2 Related work

In what follows, we review topic modeling and NLP research related to COVID-19.

4.2.1 Topic modeling

In topic modeling, a topic is often viewed as a pattern of co-occurring words that can be exploited to cluster together documents from a large collection [209]. Among methods for topic modeling we can find approaches such as the Vector Space Model (VSM) [210], Latent Semantic Indexing (LSI) [211], Probabilistic Latent Semantic Analysis (PLSA) [212] or lda2vec [54]. Related to this, one of the most well-known, standardized and widely-used methods is Latent Dirichlet Allocation (LDA) [51]. More particularly, LDA is an unsupervised clustering approach where documents can belong to multiple topics, and where each topic is a mix of words, which can be shared among topics too.

The applications of these topic modeling approaches are many and include areas such as tag recommendation [213], text categorization [214], keyword extraction [215], information filtering [216], similarity search in the fields of text mining [217], and information retrieval [218].

4.2.2 Text Mining on English COVID-19 related tweets

With the COVID-19 outbreak, different authors have tried to apply topic modeling and text mining techniques to help analyze and monitor the situation of the pandemic, with a great focus on English messages. For instance, [184] analyzed English tweets and detected the trending topics and major concerns of people with respect to COVID-19, by proposing a model based on the Universal Sentence Encoder [219]. The model first derives a semantic representation and similarity of tweets and, over those similar tweets, it applies text summarization techniques to provide a summary of different clusters. In a related line, [185] proposed a framework to analyze the topic and sentiment changes in society over time due to the COVID-19, using Twitter to collect the source data. More specifically, they used a dynamic LDA for topic modeling over fixed time intervals [220], and VADER for sentiment analysis [221]. The work [222] examined the key topics among 13.9M English tweets about COVID-19, dealing with areas such as economy and markets, spread and growth in cases, treatment and recovery, impact on the healthcare sector, and governments response. They explored the trends and variations, and how those key topics, and associated sentiments changed over a period of time of 17 weeks, between 1 January 2020 and 9 May 2020. More particularly, they used guided LDA for topic modeling [223], an LDA-variant where the model is guided to learn topics that are of specific interest, using priors in the form of seed words, and again VADER for sentiment analysis.

Also, the work [224] use LDA to detect topics such as the origin of the virus and its impact on people and countries, analyzing 2.8M English tweets. In addition, they performed sentiment analysis with `TextBlob` [225] and extracted some social network statistics for each topic, such as the number of followers, the number of likes of tweets, the number of

retweets, the user mentions, or the link sharing, calculating the interaction rate per topic. At a smaller scale (100K English tweets) and considering only the pre-crisis lockdown period (from 12 December 2019 to 9 March 2020); [226] presented a work to understand public perceptions of the trends of the COVID-19 pre-pandemic time. The analysis included time series, sentiment analysis and emotional tendency using the NRC sentiment lexicon [227], as well as topic modeling using LDA.

4.2.3 Text Mining on Spanish and Multilingual COVID-19 related tweets

As usual in NLP, most of the early efforts to monitor COVID-19 user-generated texts have focused on English. However, some work is already available for the Spanish language. For instance, [187] compare the news updates of two of the main Spanish newspapers Twitter accounts, *El País* and *El Mundo*, during the pandemic; applying topic modeling and network analysis methods. They identified eight news frames for each newspaper and split them into three clusters: the pre-crisis period (from 19 February to 14 March of 2020), the lockdown period (from 14 March to 11 May of 2020), and the recovery period (from 11 May to 3 June of 2020). Their goal was to understand how the Spanish news media covered the public health crisis on Twitter.

Besides, [188] proposed a geographical analysis of the opinion and influence of users in Twitter during the covid health crisis, considering tweets written in English and Spanish, and using LDA topic modeling. The first part of the study was a general approach to the analysis of the topics of US and UK users. The second part was an analysis of the interests of Twitter users in Spain during the confinement period (from 14 March to 22 July of 2020). To geolocate the tweets, they performed a country-level search for the English dataset, and a city or province-level search for the Spanish dataset; looking in both cases for any geographic references, both on the Twitter user location field and their biography.

The work [228] studied techniques to assess the distinctiveness of topics, key terms and features, as well as the speed of dissemination of retweets over time. They used pattern matching and topic modeling with LDA on a set of 5.5M of tweets written in multiple languages, resulting in 16 topics for English and one for Spanish, Italian, French and Portuguese, respectively. They also applied Uniform Manifold Approximation and Projection (UMAP) [229] to identify clusters of distinct topics, which discuss case spread, healthcare workers, and personal protective equipment issues.

Beyond Twitter, [186] have exploited 22K Facebook posts to track the evolution of COVID-19 related trends, with a multilingual dataset that covers seven languages (English, Arabic, Spanish, Italian, German, French and Japanese). They applied an end-to-end analytic process for discovering language-dependent topics covering the duration of the pre-crisis period and part of lockdown (from 1 January to 15 May of 2020). The experiments showed that the extracted topics corresponded to the chronological development of what has been happening, and the measures that were taken in various countries.

4.3 Methods

In what follows, we describe the methodology of our work, decomposed into two steps: (i) the preprocessing of the collected corpus, which was presented and detailed in subsection 3.2.1, and (ii) the topic modeling approach and its analysis, clustering tweets into topics and extracting representative keywords and sentences.

4.3.1 Preprocessing

We first proceed to lowercase the tweets and remove retweets. We also delete the keywords that were used to collect the tweets (see again §3.2.1) and other Twitter reserved words

such as 'rt', 'fav', 'vía', 'nofollow', 'twitter', 'href' or 'rel'. Moreover, we removed stopwords, non-words (i.e., words compounded with characters that are not alphabet letters), URLs, numbers and punctuation marks. To do this, we used spaCy[2] to tokenize the words, and the Spanish and English stopwords lists from three libraries: NLTK,[3] stop-words,[4] and stopwordsiso.[5] Besides, in order to remove extra noise and cluster more clean topics, we only kept *content words* (i.e., nouns, verbs, adjectives, and adverbs).

Finally, to reduce word sparsity we used a custom lemmatizer[6] for Spanish, which applies a rule-based lemmatization with spaCy, and relies on Wiktionary,[7] which is a collaborative free-content multilingual dictionary. After the lemmatization step, the tweets whose length is less than three characters were removed. As traditional topic modeling approaches such as LDA, based on bag-of-words, suffer if many outliers are present (which happens in NLP due to Zipf's law), we ignore terms that have a corpus frequency strictly less than three.

4.3.2 Topic modeling

For a more clear and comprehensive topic modeling analysis, we cluster the tweets in four weeks per month, except for the year 2019 (for which we collect the few tweets discussing coronavirus topics at that time), and the month of January 2020, which covers the first fortnight and not a week.

More particularly, we cluster the time of analysis into three phases. First, a pre-crisis phase, which includes tweets up to 24 January of 2020; when there were still few cases reported outside China. Second, we consider the outbreak phase, which we will consider to range from 25 January to 14 March of 2020; when the disease started to widely spread across Europe and the rest of the world, but Spain still was not under confinement. This is the period of time where the pandemic information, epidemic back then, was reported but was still not formally considered an alarm by the Spain government. Third, we cover about a month of the official lockdown period of the first wave (from 14 March to 20 April of 2020), when the Spanish government approved strict social confinement.

As introduced previously, for topic modeling, we will be using *Latent Dirichlet Allocation* (LDA)[8] with collapsed Gibbs sampling inference [230]; which processes raw text data in an unsupervised fashion to cluster documents that discuss the same topic. We chose LDA because it is standard and has proved robust for many tasks (see also §4.2.2 and §4.2.3). For each phase, we will mostly group tweets into weeks,[9] and for each week we will be extracting 10 topics. On the one hand, our goal was to facilitate comprehension and interpretability. On the other hand, it is worth noting that selecting too few topics would make the clusters very generic and unspecific, while choosing too many could make them too sparse, not representative, and hard to analyze qualitative [231]. Yet, we explored what would be in theory an optimal number of topics for different weeks using three methods: (i) the KL divergence (Arun et al., 2010) [232], (ii) the pair-wise cosine distance (Cao et al., 2009) [233], (iii) and the loglikelihood. In all cases, the results returned that the ideal number was between 5 and 20 in most cases. For example, Figure 4.1 shows the scores obtained for the three different metrics for the period that goes from 9 to 16 March 2020 (the week of the lockdown), considering from 5 to 45 topics.

[2] `https://spacy.io/usage/v2-2` and the `es_core_news_md` language model.
[3] `http://www.nltk.org/nltk_data/`
[4] `https://pypi.org/project/stop-words/`
[5] `https://pypi.org/project/stopwordsiso/`
[6] `https://github.com/pablodms/spacy-spanish-lemmatizer`
[7] `https://www.wiktionary.org/`
[8] In particular, we rely on the `https://github.com/lda-project/lda` implementation.
[9] As introduced before, we use week here in an informal sense, referring to periods of time of 7 days, but not necessarily from Monday to Sunday.

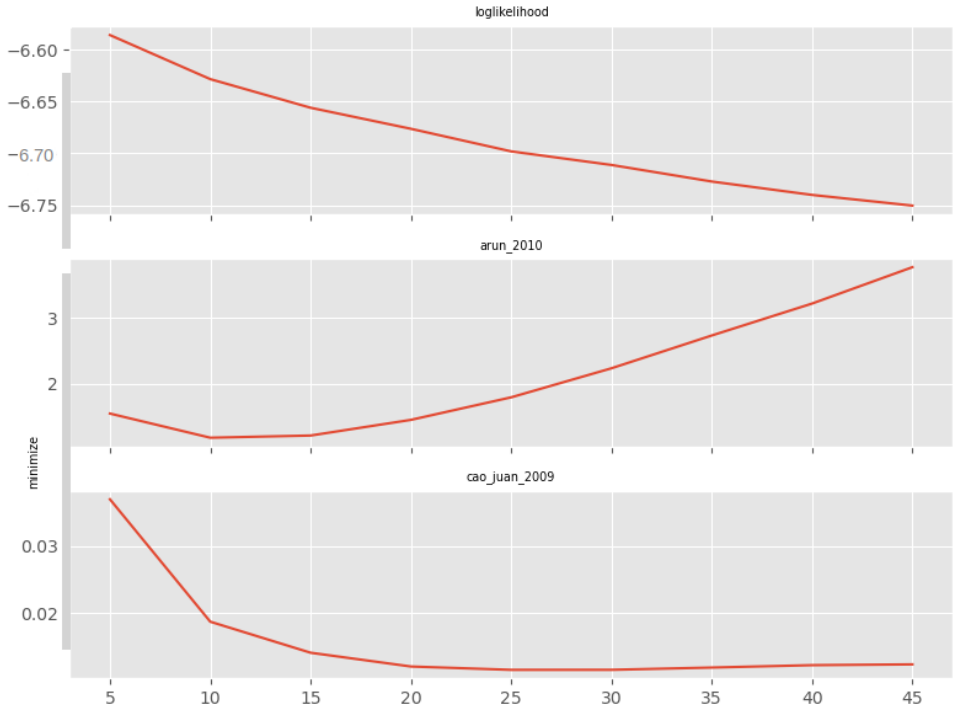


Figure 4.1: Topic model evaluation obtained for the different metrics for the period that goes from 9 to 16 March 2020, considering 5 to 45 topics. Metric direction for optimization: *maximize* for loglikelihood, and *minimize* for KL divergence - arun_2010, and pair-wise cosine distance - cao_juan_2009.

LDA setup We sampled up to 1500 epochs, and we kept the rest of the parameters to the default value in the LDA library we used, i.e., $\alpha : 0.1$, $\eta : 0.01$, where the first corresponds to the Dirichlet parameter for the distribution over topics, and the second to the Dirichlet parameter for the distribution over words. We perform the training in the *Marenostrum 4*[10] (with 24 cores), which is a High-Performance Computing (HPC) cluster.

4.3.3 Extracting top topic keywords and sentences

To extract the most representative keywords for each topic, we considered both generative (GS) (Equation 4.1) and discriminative (DS) (Equation 4.2) approaches:

$$\text{GS(w,z)} = P(w|z) \tag{4.1}$$

$$\text{DS}(w, z) = P(w|z)/[\max_{z' \neq z} P(w|z')] \tag{4.2}$$

where w represents a given word and z the topic at hand. In essence, the generative score allows extracting the words that are most representative for each topic independently, in a way that a given word could be relevant for one or more topics, potentially making such topics harder to differentiate among them. On the contrary, the discriminative score allows representing a topic by a set of keywords that are very representative for such a topic but have little relevance for the remaining ones.

Although the top keywords for each topic are useful, they might provide a limited view of what is actually being discussed. To counteract this, we also defined a generative (Equation

[10] `https://www.bsc.es/marenostrum/marenostrum`

4.3) and discriminative (Equation 4.4) routes to extract the most representative sentences (tweets) for each topic, ideally being able to determine the topic by simply reading a few documents. The motivation to define these two different routes is the same than the one we made to extract the top keywords.

$$\mathrm{GS}_{\mathrm{sent}}(s, z) = \sum_{w \in s} \mathrm{GS}(w, z)/\mathrm{Length}(s) \tag{4.3}$$

$$\mathrm{DS}_{\mathrm{sent}}(s, z) = \sum_{w \in s} \mathrm{DS}(w, z)/\mathrm{Length}(s) \tag{4.4}$$

where s is the input document, for which we consider its length, in order not to only select the longest documents; although in the case of Twitter this is less of an issue than in other topic modeling approaches that must deal with actual long documents.

The full code is available.[11]

Topic	Discriminative Keywords	Generative Keywords
'W1' (from January to December of 2019)		
2	respiratorio, enfermedad, gripe	respiratorio, gripe, enfermedad
Magnifica guia para diferenciar los sintomas que causa la gripe y otros virus respiratorios. Junto con la gripe siguen circulando rinovirus, virus respiratorio sincitial y coronavirus, entre otros. <URL>		
1	enfermedad, gripe, respiratorio	enfermedad, respiratorio, gripe
@user informa de 27 casos de neumonia atipica, probablemente virica, en Wuhan (Hubei, China) en fecha 31/12/2019. El SARS (coronavirus) se inicio asi en 2003. Habra que seguir evolucion y esperar el diagnostico. <URL>		
W2-3 (from 1 to 16 January of 2020)		
8	alerta, hospital, poner, red, oms, china, mundial, mundo	china, oms, alerta, hospital, poner, mundial, mundo, red
UN NUEVO CORONAVIRUS PONE EN ALERTA A CHINA <URL>vía @user		
5	confirmar, japón, chino, infección, caso, china, animal, aparición	caso, confirmar, japón, china, infección, chino, ciudad, identificar
Japón confirma el primer caso de coronavirus vía @user <URL>		
W4 (from 17 to 24 January of 2020)		
9	emergencia, declaración, declarar, organización, reunión, convocar, decisión, determinar	oms, emergencia, internacional, declarar, mundial, alerta, salud, china
La OMS no declaró la emergencia por el coronavirus <URL>		
1	millón, cuarentena, habitante, frenar, ampliar, pekín, transporte, aislar	china, ciudad, wuhan, millón, cuarentena, persona, cerrar, brote
Más de once millones de chinos, en cuarentena por el coronavirus <URL>		

Table 4.1: Some representative topics for the weeks corresponding to the **pre-crisis** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

Limitations Sociolinguistic studies that collect data from social media such as Twitter can suffer from biases that can be hard to measure, identify or correct. For instance, it is well-known that a small percentage of Twitter users generate the majority of content [234]. In this line, we believe that many of the collected tweets have their origin in newspapers and journalists' accounts, which conditions how other users tweet about this topic on Twitter, and therefore the detected topics can be heavily dependent on how national media decide to spread the news. Yet, this is the natural behavior of this network, and in this particular work, we decided not to control for this variable.

[11] `https://github.com/mmaguero/twitter-analysis`

4.4 Results

We consider sixteen sets of tweets (mostly grouped on a weekly basis), extracting the ten most representative topics for each one according to LDA. To refer to the topics, we will represent them with the top eight keywords and the most salient tweets. For clarity, and due to a large number of weeks and topics, we will just illustrate and analyze some relevant topics extracted by our approach for different weeks, and try not to repeat common topics that span through the whole period. Usernames and URLs are cut due to anonymityreasons.

4.4.1 Pre-crisis time

During this pre-crisis time, it is possible to see how the model captures that the COVID-19 was still not a concern for the Spanish society, which perceived the disease as an external problem, as reflected in many of the extracted topics. For clarity, Table 4.1 illustrates some relevant topics with top keywords and tweets, but we briefly discuss the content of the table below. To assess the relevance of the topics, we will be matching those against news from the newspapers that were published at the time in different Spanish media.

'W1' (from January to December of 2019) For the year 2019, we only could extract a total of seven topics, since the corresponding subset of tweets related to COVID-19 or coronavirus was still tiny (a total of 43 tweets after preprocessing). Still, we believe the results are interesting, since we observed that at this time most the Spanish tweets dealing with coronavirus still had to do with veterinarian diseases or even the zoonosis of coronavirus (i.e., how it is transmitted between animals and humans through the air), Yet, we found a few relevant tweets about COVID-19 that started to show up. We illustrate this as part of Table 4.1.

W2-3 (from 1 to 16 January of 2020) This time can be considered as the start of the emergency [235]. In this line, we observed how our model started to identify this situation as well, clustering tweets about the World Health Organization (WHO) alerts to hospitals about symptoms, procedures, etc., and also about the increase in the number of cases in China.

W4 (from 17 to 24 January 2020) The crisis started to expand and from our model we see how the topics differ from previous weeks (see the third group of rows in Table 4.1). For instance, it shows how China started to apply restrictions in many locations of its territory (e.g., Wuhan) [236].

4.4.2 Outbreak time

In this phase, we see how the LDA approach reflects emergency declarations, the first cancellations of massive events in Spain, as well as the first suspicious cases; causing, in consequence, an increase of the concern among the Spanish society, which started to look and ask for sanitary products. This is also the phase where the approach captures a transition from international to national concerns. We will breakdown this more in detail in the next paragraphs, matching again the topics against news from the newspapers to qualitatively verify the quality of the extracted topics. Table 4.2 illustrates such topics with the top keywords and tweets from the model.

W5 (from 25 to 31 January of 2020) During this week, the approach kept identifying online discussions about the WHO emergency declarations, considering COVID-19 as a global coronavirus threat [237]. Also, the approach extracted topics related to international

Topic	Discriminative Keywords	Generative Keywords
W5 (from 25 to 31 January of 2020)		
9	oms, emergencia, declarar, declaración, sanitaria, organización, comité, convocar	oms, internacional, emergencia, salud, declarar, alerta, mundial, china
Declara OMS emergencia por coronavirus - Vía @user <URL>		
1	vuelo, cerrar, suspender, frontera, kong, hong, rusia, aerolínea	china, vuelo, cerrar, suspender, brote, frontera, evitar, kong
Iberia suspende los vuelos a Shanghái por el coronavirus <URL>...		
W6 (from 1 to 7 February of 2020)		
4	alertar, acusar, news, silenciar, intentar, bbc, difundir, confusión	médico, china, chino, morir, alertar, wuhan, muerte, wenliang
Por favor lean. Porque esto no lo va a contar ningún medio que alerte sobre el coronavirus . <URL>...		
1	gomera, alemán, ingresado, contacto, jalisco, vitoria, ecuador, isla	caso, españa, gomera, paciente, sospechoso, hospital, salud, síntoma
En España ya tenemos un caso de coronavirus ,en La Gomera ,un alemán.		
W7 (from 8 to 14 February of 2020)		
6	mallorca, negativo, británico, ingresado, palma, princess, diamond, gomera	caso, mallorca, españa, crucero, paciente, confirmar, sospechoso, salud
Confirman un caso de coronavirus en Palma de Mallorca <URL>... <URL>		
3	sony, amazon, gsma, bajas, lg, nvidia, ericsson, intel	mobile, congress, barcelona, mwc, empresa, cancela, cancelar, sony
Tras las bajas de LG, Ericsson, NVidia, Amazon y Sony #coronavirus #MWC2020 <URL>...		
W8 (from 15 to 21 February of 2020)		
5	crucero, diamond, princess, pasajero, colombiano, camboya, evacuado, ucrania	crucero, cuarentena, japón, caso, diamond, princess, pasajero, wuhan
NUEVOS CASOS DE CORONAVIRUS EN CRUCERO DIAMOND <URL>... <URL>		
6	mobile, barcelona, cancelación, cancelar, maratón, evento, mwc, congress	mobile, china, tokio, barcelona, cancelar, cancelación, maratón, guerra
Suspenden el Mobile World Congress de Barcelona por el coronavirus <URL>... <URL>		
W9 (from 22 to 29 February of 2020)		
6	mano, farmacia, lavarse, gel, desinfectante, alcohol, agotar, carne	mascarillas, mano, gente, mascarilla, evitar, comprar, miedo, hospital
Cómo prevenir el #coronavirus . Lávate las manos, lávate las manos, lávate las manos..... lávate las manos. <URL>... <URL>		
1	bolsa, economía, mercado, caída, ibex, pérdida, crecimiento, wall	china, bolsa, economía, mercado, crisis, mundial, impacto, económico
'Esto es mercado. Esto me pone' @user #bolsa #COVID19 <URL>...		
W10 (from 1 to 8 March of 2020)		
7	mano, metro, lavarse, higiene, agua, gel, jabón, lavar	mano, evitar, medido, contagio, mascarillas, covid, persona, tomar
me voy a lavar las manos que no quiero el coronavirus		
4	patología, contagioso, anciano, letalidad, diferencia, estacional, comparación, hambre	gripe, persona, año, morir, mortalidad, gente, matar, enfermedad
Se llama Virus Corona Patologías Previas		

Table 4.2: Some representative topics for the weeks corresponding to the **outbreak** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

restrictions, such as the airplane company Iberia suspending flights to Shanghai [238], at the same time that Russia closed its frontiers with China [239].

W6 (from 1 to 7 February of 2020) Following the trend of announcing emergency declarations, the model started to identify international issues, such as the infection and posterior death of Li Wenliang [240], a Chinese doctor that alerted about the first cases of COVID-19 in December 2019, but also national ones; such as the confirmation of the first case of coronavirus in Spain, in the Canary Island of La Gomera [241]. This matches the time where the number of cases seemed to start to spread (still slowly) all around the world.

W7 (from 8 to 14 February of 2020) During this week, the coronavirus started to have an important economic effect in Spain, which is reflected by the model, discovering

topics that showed how users discussed the potential (finally confirmed during this week too) cancellation of the 2020 Mobile World Congress (MWC 2020), which usually takes place in Barcelona [242]. On the healthcare side, additional (few) cases started to be reported in Spain, such as in Mallorca, where was reported the second Spanish case of COVID-19 [243]. During this and the next weeks, we started to observe how there is a slow transition from international to national topics.

W8 (from 15 to 21 February of 2020) During this week, the topics were in line with those discussed in the previous weeks, such as the cancellation of the MWC 2020 (see Table 4.2) and its repercussion. These 'last-long' topics made sense at the time, since the cancellation of the MWC 2020 was the first massive event cancelled in Spain, with important economic consequences. Other international issues such as the sustained increase of cases in China or in the cruise ship Diamond Princess [244] seemed to occupy Twitter users during this time, too.

W9 (from 22 to 29 February of 2020) These are the final days before the lockdown period, and in retrospect, it is easy to see how some of the topics extracted reflected the immediate seriousness of the situation. We see how the model captures that the WHO advised the public [245] to wash hands frequently. It is interesting to see in Table 4.2 how 'farmacia' (pharmacy) appears together with 'gel' (gel), 'lavarse' (to wash), 'mano' (hand) and 'alcohol' (alcohol), 'agotar' (to run out of) among the top keywords for the corresponding topic. In this context, it is well-known that these products were scarce in pharmacies and stores, and actually, this problem lasted for a long during the lockdown period. Also, related to the immediate seriousness of the situation, the model captured how despite not being confined, the world economy started to suffer from the stocks set for the worst week since 2008 [246].

W10 (from 1 to 8 March of 2020) Just before the lockdown, we observe how among the topics extracted there are topics that we see every day in the current pandemic life. For instance, as shown in Table 4.2, we kept seeing the importance of washing hands and keeping good hygiene with the use of soap [245]. Also, 'metro' (underground) is a top keyword of such topic, since at that time there was a discussion about the chances of getting infected (e.g., in the public transport) [247]. On a different topic, we see what it seems to be a discussion comparing the flu and covid, and how they affect the population, which was a popular comparison at the time.

4.4.3 Lockdown time

During the lockdown phase (until April), we can observe in Table 4.3 how the topics discussed mostly focused on the worst consequences of the pandemic, such as the big economic crisis, the large number of deaths per day, and also some collective actions such as thanking the healthcare workers. Again, we give a brief explanation below these lines, and match the topics against news in the media.

W11 (from 9 to 16 March of 2020) Here we consider the week where the Spanish society stopped having free movement. More particularly, the government approved strict social confinement on 14 March of 2020 [248]. Besides, the model found topics about the acknowledgment to the healthcare workers and the solidarity applause [249], which was very popular in Spain during the lockdown period. In a related line, topics like this one also captured the feeling of the importance of staying at home to prevent becoming infected and reduce the workload of these workers.

Topic	Discriminative Keywords	Generative Keywords
W11 (from 9 to 16 March of 2020)		
3	aplauso, frenalacurva, aplausosanitario, cuarentenaya, yoelijoserresponsable, felizlunes, arena, agradecimiento	covid, yomequedoencasa, casa, quedateencasa, cuarentena, coronavirusesp, cuarentenacoronavirus, responsabilidad
Aplausos para que suenen más que los truenos que hoy hay en Madrid. Hoy mis aplausos para todos. Para \Saldremos de esta / #COVID19 <URL>		
8	ocasionado, aprobar, pymes, paliar, erte, fiscal, boe, hipoteca	covid, medido, crisis, gobierno, alarma, situación, sanitaria, empresa
#RealDecreto 463/2020 #estadodealarma #COVID19 <URL>#pymes #Autonomo #Cordoba @user <URL>		
W12 (from 17 to 24 March of 2020)		
6	respirador, fabricar, ifema, impresora, envío, coronavirius, epis, todosobremovil	covid, hospital, sanitario, mascarillas, madrid, personal, estevirusloparamosunidos, quedateencasa
#ElonMusk puede que empiece a fabricar respiradores #COVID19 <URL>		
1	higiene, jabón, distanciamiento, acatar, lavado, fanb, geacam, comerciales	covid, medido, evitar, contagio, prevención, propagación, salud, tomar
Entre más higiene se tenga, mayor es la protección ante los patógenos como el #COVID19 <URL>...		
W13 (from 25 to 31 March of 2020)		
1	erte, pago, despido, prestación, contrato, fiscal, ertes, alquiler	crisis, medido, gobierno, empresa, económico, trabajador, autónomo, sanitaria
Información para los afectados por ERTE debido al COVID19 . #ERTE #Coronavirus <URL ><URL>		
3	civil, guardia, desinfección, desinfectar, higiene, cumplimiento, jabón, estación	medido, persona, evitar, contagio, salud, seguridad, prevención, casa
Unos 400 guardias civiles con coronavirus en #CLM , según la @user @user -. Vía @user <URL>... <URL>		
W14 (from 1 to 7 April of 2020)		
1	animal, respiratorio, tigre, gato, mascota, zoo, bronx, contaminación	persona, paciente, enfermedad, síntoma, casa, contagio, evitar, matar
Si los tigres se contagian de coronavirus , ¡ojito los que tenéis gato!		
8	confirmado, cifra, elevar, defunción, diarios, ascender, activos, descender	caso, fallecido, españa, muerte, muerto, dato, número, país
637 muertes por coronavirus en un día, la cifra más baja en 13 días <URL>		
W15 (from 8 to 14 April of 2020)		
2	cifra, curado, reino, récord, ascender, contabilizar, diagnosticado, acumular	caso, fallecido, españa, muerto, muerte, dato, número, persona
Las 510 muertes por COVID-19 en un día, la cifra más baja desde el 23 de marzo <URL>		
5	johnson, intensivo, boris, testimonio, alta, universitario, clmpressdigital, sosprisiones	hospital, médico, paciente, sanitario, madrid, personal, persona, profesional
Coronavirus : Boris Johnson fue dado de alta. <URL>		
W16 (from 15 to 20 April of 2020)		
5	luis, sepúlveda, escritor, homenaje, chileno, fútbol, club, dep	año, morir, hospital, luis, fallecer, quedateencasa, yomequedoencasa, historia
Luis Sepúlveda muere por coronavirus <URL>... <URL>		
2	distanciamiento, prohibidorendirse, enestafamilianadieluchasolo, yonosoyungastosuperfluo, bicicleta, espandemia, comunidadvalenciana, saltarse	confinamiento, quedateencasa, yomequedoencasa, cuarentena, medido, casa, evitar, alarma
¿El distanciamiento social podría ir incluso más allá de 2021? #COVID19 #coronavirus <URL>		

Table 4.3: Some representative topics for the weeks corresponding to the **lockdown** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

W12 (from 17 to 24 March of 2020) For this week, the model extracted topics discussing the personal hygiene measures to combat the COVID-19. The topics also reflect the lack of equipment in the hospitals, which was a problem at the beginning of the pandemic. More particularly, the model was able to identify as a topic the lack of ventilators in Spain, and also the rest of the world, as reflected by the most salient discriminative tweet. This matches the news at the time, which discussed the use of 3D printers to provide such ventilators [250], or hacking some objects to adapt them for medical use [251].

W13 (from 25 to 31 March of 2020) This week covers the last days of March 2020. Due to the strict confinement, topics concerning job losses and the measures taken by the government to counteract the situation (e.g. the so-called ERTEs) started to arise [252, 253]. Among the rest of the topics of this week, we also would like to remark the massive infection of public workers, such as the Guardia Civil officers in Castilla La-Mancha [254]. The infection of public workers during this time of the pandemic was also widely discussed in the news [255].

W14 (from 1 to 7 April of 2020) On the national side, some topics reflected the number of casualties per day. More particularly, the beginning of April corresponded to the peak of the first wave, and the beginning of the decreasing trend in the number of infections and deaths per day [256]. A bit on a different line, we found topics discussing more diverse aspects of COVID-19, such as the infection in the Zoo of Bronx (New York, USA) of tigers and lions [257].

W15 (from 8 to 14 April of 2020) Here, we would like to remark a topic related to an international breaking news, and more particularly about Boris Johnson (the UK Prime Minister) being infected by the coronavirus, together with his evolution, when he even entered the ICU [258]. On the national side, the models kept detecting topics related to the number of deaths in Spain, which was still high and dynamic during that time, but reached some local minima these days [259].

W16 (from 15 to 20 April of 2020) For the last days of our study, the model found relevant topics too, such as the death of the Chilean writer Luis Sepúlveda [260] due to COVID-19, or topics related to the need of keeping social distancing, maybe even for months [261], as reflected by some of the most representative tweets.

4.4.4 Quantitative evaluation

We performed a small human evaluation to quantitatively estimate the quality of the extracted topics. We took 20 topics randomly from all periods. Then, two annotators were in charge of: (i) determining if given the top 8 keywords and 3 top sentences made possible to infer a topic, (ii) determining if for each top topic word (according to the discriminative score) they belonged to the inferred topic, and (iii) the same as in (ii), but for the 3 most representative sentences. See section B.2 for more details about the process. We calculated the percentage of times both annotators positively labelled a sample, obtaining scores of 80%, 56.88% and 71.66% for (i), (ii), and (iii), respectively. In addition, we calculated (ii) but taking into account only the first 3 top keywords of the topic, yielding a score of 75% of positive samples.

4.5 Conclusion

This chapter used a topic modeling approach to shed some light on the topics discussed in Spain during the early stages of the COVID-19 pandemic, including a period of pre-crisis, the outbreak of the disease, and the beginning of the confinement. We used the tweets collected in subsection 3.2.1, which were cleaned them to keep only Spanish tweets that were written in Spain. On this collection of tweets, we applied a Latent Dirichlet Allocation model that learned to cluster such tweets according to the topic they discuss. To represent the topics, we used generative and discriminative routes to extract the most salient keywords and sentences. To verify the quality of the extracted topics, we performed a qualitative analysis matching the topics against relevant news in the newspapers at the

same period of time, and a small quantitative evaluation. Overall, the topics show that during the pre-crisis period, users focused on the international panorama than the local situation, while during the outbreak and lockdown phases they focused the most on the Spanish emergency, considering health and economic problems.

Acknowledgements

We are grateful to the Barcelona Supercomputing Center (BSC) through the Spanish Plan for advancement of Language Technologies 'Plan TL' and the *Secretaría de Estado de Digitalización e Inteligencia Artificial* (SEDIA), for giving us access to the hardware necessary to perform the training of the LDA models.

Chapter 5

Sentiment Analysis in Low-resourced Code-switched Languages: The Case of Guarani and Jopara [1]

In this chapter we explore sentiment analysis in indigenous low-resourced code-switched languages. More particularly, we perform experimentation with the Guarani and Jopara languages with the dataset constructed in subsection 3.2.2.

5.1 Introduction

Indigenous languages have been often marginalized, an issue that is reflected when it comes to design natural language processing (NLP) applications, where they have been barely studied [262]. One of the places where this is greatly noticed is Latin America, where the dominant languages (Spanish and Portuguese) coexist together with hundreds of indigenous languages such as Guarani, Quechua, Nahuatl or Aymara.

In this context, the Guarani language plays a particular role. It is an official language in Paraguay, Bolivia and Corrientes Province (Argentina) and it is also an official language of Mercosur,[2] a common market organization that involves a number of South American countries. Besides, it is spoken in other regions, e.g., some localities of Mato Grosso do Sul (Brazil), alongside with their official languages. Overall, it has about 8M speakers. Its coexistence with other languages, mostly Spanish, has contributed to its use in code-switching setups [263, 264, 265] and led to Jopara, a code-switching between Guarani and Spanish [266], with flavours of Portuguese and English.

Despite its official status, there are still few NLP resources developed for Guarani and Jopara. In [267], authors developed a parallel Spanish-English-Guarani corpus for machine translation. Similarly, authors of [268] developed a Guarani-Spanish parallel corpus aligned at sentence-level. There are also a few online dictionaries and translators from Guarani to Spanish and other languages.[3] Beyond machine translation, in the work [270] was released a corpus for Guarani speech recognition that was collected from the web; and another

[1] Chapter based mainly on the content of a paper published in *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* (co-located in *2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*) [96] (see section A.5).

[2] `https://www.mercosur.int/en/`

[3] `https://gn.wiktionary.org/`, `https://es.duolingo.com/dictionary/Guarani/`, `https://www.paraguay.gov.py/traductor-guarani`, `https://www.iguarani.com/`, `https://glosbe.com/gn`, and `Mainumby` [269].

work [271] presented a system for cross-lingual word sense disambiguation from Spanish to Guarani and Quechua languages. There also are a few resources for Part-of-Speech-tagging and morphological analysis of Guarani, such as the work by [272] and Apertium;[4] and also for parsing, more specifically for the Mbyá Guarani variety [273, 274], under the Universal Dependencies framework.

In the context of sentiment analysis (SA) [15, 112], and more particularly classifying the polarity of a text as positive, negative or neutral, we are not aware of any previous work; with the exception of [40]. They presented a sentiment corpus for the Paraguayan Spanish dialect, which also includes words in English and Portuguese. However, there were few, albeit relevant, words of Guarani (70) and Jopara[5] (10), in comparison to the amount of the ones in Spanish (3,802) [40, p. 40, Table II]. Overall, SA has focused on rich-resource languages for which data is easy to find, even when it comes to code-switching setups [275], maybe with a few exceptions such as English code-switched with languages found in India [168, 276, 277]. In this context, although some previous work has developed multilingual lexicons and methods [278, 279]; for languages such as Guarani and other low-resource cases (where web text is scarce), it is hard to develop NLP corpora and systems.

Contribution We train a set of neural encoders and also traditional machine learning models for polarity classification with the presented dataset in JOSA (see subsection 3.2.2), in order to have a better understanding of how old versus new models perform in this low-resource setup, where the number of data matters.

5.2 Methods

Due to the low-resource setup, we run neural models and pre-trained language models, but also other machine learning models, such as complement naïve Bayes (CNB) and Support Vector Machines (SVMs) [280], since they are less data-hungry, and could help shed some light about the real effectiveness of neural models on Jopara texts. We will report results both on the JOSA unbalanced and balanced setups (see subsection 3.2.2). In all cases, the selection of the hyperparameters was done over a small grid search based on the dev set. We report the details in Appendix C.

Naïve Bayes and SVMs We tokenized the tweets[6] and represented them as a 1-hot vector of uni-grams with a TF-IDF weighting scheme. We used [281] for training.

Neural networks for text classification We took into account neural networks that process input tweets as a sequence of token vector representations. More particularly, we consider both long short-term memory networks (LSTM) [143] and convolutional neural networks (CNN) [282], as implemented in NCRF++ [283]. Although the former is usually more common in many NLP tasks, the latter has also shown traditionally a good performance on sentiment analysis [284].

For the input word embeddings, we tested: (i) randomly initialized word vectors, following an uniform distribution, (ii) and pre-trained non-contextualized representations and more particularly, FastText's word vectors [69] and BPEmb's subword vectors (including the multilingual version, which supports Guarani) [174]. In both cases, we also concatenate a second word embedding, computed through a char-LSTM (or CNN).

[4] `https://github.com/apertium/apertium-grn`

[5] Tokens that mix n-grams of characters from Guarani and Spanish, e.g.: '*I understand*' would be '*entiendo*' (es), '*ahechakuaa*' (gn) and '*aentende*' (jopara).

[6] We used the TweetTokenizer from the NLTK library.

Pre-trained language models We also fine-tuned recent contextualized language models on the JOSA training set. We tested BERT [12] including: (i) beto-base-uncased (a Spanish BERT) [285], (ii) multilingual bert-base-uncased (mBERT-base-uncased, pre-trained on 102 languages), and (iii) and the original bert-base-uncased (pre-trained on English). The idea for testing original BERT with JOSA, is due to that Jopara may have words borrowed from English (e.g., in the fields of IT, fashion, sport, etc., as is commonly the case for other languages and dialects) [286] or the use of Netspeak [287].[7] We also tried more recent variants of multilingual BERT, in particular, (iv) XLM [38]. Note that BERT models use a wordpiece tokenizer [288] to generate a vocabulary of the most common subword pieces, rather than the full tokens, and that in the case of the multilingual models, none of the language models used considered Guarani during pre-training.

5.3 Experiments and results

Reproducibility The baselines[8] are available at `https://github.com/mmaguero/josa-corpus`.

We run experiments for the unbalanced and balanced versions of Jopara Sentiment Analysis dataset (JOSA) (see subsection 3.2.2), evaluating the macro-accuracy (to mitigate the impact of the neutral class in the unbalanced setup). Table 5.1 shows the comparison. Note that all models, even the non-deep-learning models, only use raw word inputs and do not consider any additional information or hand-crafted features,[9] yet they obtained results that are in line with those of more recent approaches.

With respect to the experiments with CNNs and BiLSTMs encoders, we tested different combinations using character representations, which output is first concatenated to a second external word vector (as explained in §5.2), and then fed to the encoder. Among those, the model that used a character-level CNN and a word-level BiLSTM encoder obtained the best results. This finding are in line with are reviewed in chapter 2. Still, the difference with respect to traditional machine learning models is small. We hypothesize this might be due to the low-resource nature of the task. Finally, the pre-trained language models that use transformers architectures, in particular BETO, obtain overall the best results, despite not being pre-trained on Guarani. We believe this is partly due to the presence of Spanish words in the corpora and also to the cross-lingual abilities that BERT model might explode, independently of the amount of word overlap [289].

Error analysis on the balanced version of JOSA Figure 5.1 shows the confusion matrices for a representative model of each machine learning family (based on the accuracy): (i) CNB, (ii) the best BiLSTM-based model (CNN-BiLSTM), and (iii) Spanish BERT (BETO). There seem to be different tendencies in the miss-classifications that different models make. For instance, CNB tends to over-classify tweets as negative, while both deep learning models show a more controlled behaviour when predicting this class. Although for the three models neutral tweets seem to be the easiest to identify, both deep learning models are clearly better at it. Finally, when it comes to identifying positive tweets, BETO seems to show the overall best performance. These different tendencies indicate that an ensemble method could be beneficial for low-resource setups such as the ones that JOSA represent, since the models seem to be complementary to a certain extent. In this context, we would like to explore

[7] Netspeak is the Internet slang, an unofficial form of language used by people on the social media to communicate to one another (e.g., 'LOL' meaning 'laugh out loud').

[8] Contact the authors for more details.

[9] In order to keep a homogeneous evaluation setup.

Model	Corpus: Unbalanced	Corpus: Balanced
Traditional Machine Learning		
CNB	0.50	0.55
SVM	0.55	0.54
Neural networks for text classification		
^{C}CNN-^{W}BiLSTM	0.45	0.57
^{C}CNN-^{W}CNN	0.46	0.52
^{C}BiLSTM-^{W}CNN	0.49	0.53
^{C}BiLSTM-^{W}BiLSTM	0.45	0.53
$_{BPEmb,gn}$ ^{C}CNN-^{W}BiLSTM	0.46	0.53
$_{BPEmb,gn}$ ^{C}CNN-^{W}CNN	0.42	0.49
$_{BPEmb,gn}$ ^{C}BiLSTM-^{W}CNN	0.42	0.50
$_{BPEmb,gn}$ ^{C}BiLSTM-^{W}BiLSTM	0.44	0.52
$_{BPEmb,es}$ ^{C}CNN-^{W}BiLSTM	0.45	0.52
$_{BPEmb,es}$ ^{C}CNN-^{W}CNN	0.42	0.50
$_{BPEmb,es}$ ^{C}BiLSTM-^{W}CNN	0.45	0.50
$_{BPEmb,es}$ ^{C}BiLSTM-^{W}BiLSTM	0.46	0.49
$_{BPEmb,m}$ ^{C}CNN-^{W}BiLSTM	0.47	0.52
$_{BPEmb,m}$ ^{C}CNN-^{W}CNN	0.42	0.50
$_{BPEmb,m}$ ^{C}BiLSTM-^{W}CNN	0.43	0.48
$_{BPEmb,m}$ ^{C}BiLSTM-^{W}BiLSTM	0.46	0.50
$_{FastText,gn}$ ^{C}CNN-^{W}BiLSTM	0.46	0.53
$_{FastText,gn}$ ^{C}CNN-^{W}CNN	0.44	0.48
$_{FastText,gn}$ ^{C}BiLSTM-^{W}CNN	0.42	0.51
$_{FastText,gn}$ ^{C}BiLSTM-^{W}BiLSTM	0.43	0.49
$_{FastText,es}$ ^{C}CNN-^{W}BiLSTM	0.46	0.52
$_{FastText,es}$ ^{C}CNN-^{W}CNN	0.44	0.51
$_{FastText,es}$ ^{C}BiLSTM-^{W}CNN	0.46	0.46
$_{FastText,es}$ ^{C}BiLSTM-^{W}BiLSTM	0.46	0.49
Pre-trained language models		
BETO$_{base,uncased}$	**0.64**	**0.64**
mBERT$_{base,uncased}$	0.55	0.58
BERT$_{base,uncased}$	0.56	0.58
XLM-MLM-TLM-XNLI-15	0.46	0.49

^{C}Encodes character sequence. ^{W}Encodes word sequence.

Pre-trained embeddings are represented with a prefix together with their language ISO 639-1 code (except for m: multilingual).

Table 5.1: Experimental results on JOSA, both on the balanced and unbalanced setups. Macro-accuracy reported.

this line of work in the future, following previous studies such as [165], which showed the benefits of combining different machine learning models for Hindi-English code-switching SA.

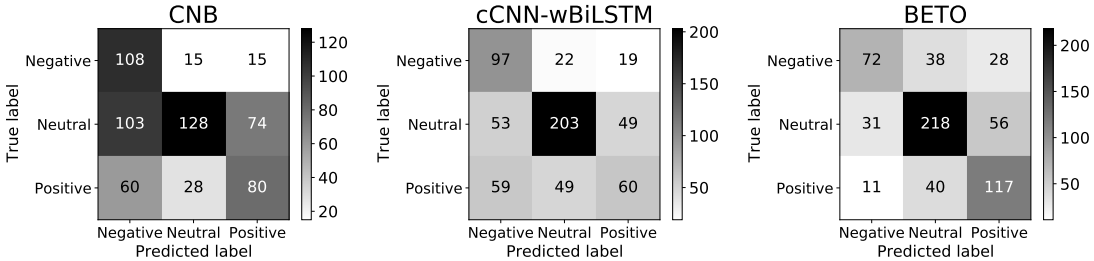


Figure 5.1: Confusion matrix for the balanced version of JOSA and the predictions of a representative member of each machine learning family: CNB, a BiLSTM-based model and Spanish BERT (BETO).

5.4 Conclusion

This chapter explored sentiment analysis on Jopara, a code-switching language that mixes Guarani and Spanish. We built several machine learning (naïve Bayes, SVMs) and deep learning models (BiLSTMs, CNNs and BERT-based models) to shed light on how they perform on this particular low-resource setup. Overall, transformers models obtain the best results, even if they did not consider Guarani during pre-training. This poses interesting questions for future work such as how cross-lingual BERT abilities [289] can be exploited for this kind of setup, but also how to improve language-specific techniques that can help process low-resource languages efficiently.

Acknowledgements

We would like to thank Yliana V. Rodríguez (Leiden University Centre for Linguistics - LUCL) and Josefina Bittar Prieto (Languages and Applied Linguistics - University of California, Santa Cruz) for fruitful discussions on the border between Guarani and Jopara.

Chapter 6

Affective Computing in Guarani and Jopara: Evaluating Guarani BERT Representations [1]

Continuing with the experimentation carried out in chapter 5, in this case we explore the affective component. More particularly, we perform Affective Computing in Guarani/Jopara with the dataset constructed in subsection 3.2.3 and propose a set of Guarani BERT Representations.

6.1 Introduction

As mentioned in the chapter 5, in Latin America, hundreds of (endangered) indigenous languages, such as Guarani, Quechua, Nahuatl or Aymara, coexist with the dominant ones, mainly Spanish and Portuguese. In some cases (such as for instance Guarani or Nahuatl), the use of indigenous and colonizing languages is deeply intertwined. Yet, although in the context of natural language processing (NLP), developing and democratizing technologies for indigenous languages is arising increasing interest in the last years, the outcomes are still limited [262, 290, 291].

The motivation to develop resources and models for these types of languages is determined by a number of factors, such as: (i) the great difficulty of preserving these languages in the digital world if applications are not developed and their digitization is not on par with that of privileged languages, (ii) the risk of the native speakers abandoning low-resource mother languages for resource-rich ones if they cannot exploit the advantages of digital NLP applications, or (iii) avoiding biases in social-linguistic analysis that mostly only consider privileged languages; among others.

For the particular case of this chapter, again we will focus on Guarani. Guarani (`gn`) is a language belonging to the Tupi-Guarani family and it has about 8M speakers. Among the indigenous languages, Guarani has a relative status of privilege: it is an official language in Paraguay, Bolivia, Corrientes Province (Argentina) and some localities of Mato Grosso do Sul (Brazil), even in the Mercosur.[2] Due to its intertwined use with other languages, mainly Spanish but also with Portuguese and English, Guarani is mostly used in code-switching setups, giving rise to Jopara, a code-switching between Guarani and Spanish [266]. Table 6.1 shows a tweet written in jopara (the username is cut off due to anonymity).

In the context of automatizing subjective analysis, we are aware of little work developing affective resources for Guarani, in a similar way to those available for any other language.

[1] Chapter based on an article that is pending to submitted to a journal.

[2] `https://www.mercosur.int/en/`

@USER ha'e la persona iporaveva ko mundore ha ndaipori mava odiscutivaera upeva.
@USER is the most beautiful person in the world and there is no one who can argue with that.

Table 6.1: Tweet wrote in Jopara and its translation into English. Note that Spanish words and roots are shown in red.

More particularly, sentiment analysis tools are needed to democratize the use of the internet and social media for Guarani speakers, not only for indigenous communities but also for those who have Guarani as their mother tongue.

Contribution We present a framework for affective analysis of Guarani/Jopara. We study two types of neural models for the task at hand: (i) recurrent LSTM networks and CNN, and (ii) Transformer-based models [12, BERT], including multilingual models, a model specifically trained on Guarani texts, and Spanish BERT [285, BETO]. Note that we trained (ii) with limited resources (∼1M tokens on a single 24GB GPU), in order to test whether this approach results useful for truly low-resource scenarios, where the amount of data and resources are really scarce.

6.2 Related work

New NLP resources have recently emerged for Guarani and Jopara. For example, [292] experimented with morphological information to improve its machine-translation system for the Guarani-Spanish pair. In the first workshop on NLP for indigenous languages of the Americas [290], [293] presented a morphological analyzer for Paraguayan Guarani and, [196] experimented in a parallel Guarani-Spanish set of news articles as well as in a monolingual set of tweets. Additionally, in the same workshop, [60] proposed a machine-translation shared-task, where participants built systems that translated between Spanish and Guarani (and other nine Latin American indigenous languages). Along the same lines, [294] constructed and experimented with Guarani-Spanish machine-translation resources, on both directions of the language pair.

On the other hand, our work presented in chapter 3, describes the findings and logistical difficulties of Jopara's Sentiment Analysis (SA), in turn, how to resolve them, in a corpus collected from Guarani-dominated tweets. Along these lines, traditional methods that usually work for collecting data in resource-rich languages failed in attempts to collect them. For this reason, it is not easy to perform NLP tasks in this kind of context. Note that, in addition, Guarani is very intertwined with Spanish, resulting in extra challenges in this particular setting.

Beyond sentiment analysis (SA) [112], in the context of text-based affect detection [295], more particularly in emotion recognition [23, 296], humor detection [297] and, offensive language identification (and hate speech) [32], we are not aware of any previous work for Guarani or Jopara. In general, affect detection has been devoted to languages whose textual content is relatively easy to collect, even when code-switched settings happen [298], perhaps with some exceptions such as code-switched English with Asian languages [299, 300, 301]. In this context, even although some previous research has developed multilingual approaches [302, 303, 304]; for languages such as Guarani and other truly resource-poor languages [11, p. 7658, Table 1], it is challenging to develop new NLP corpora and systems.

6.3 Methods

In this section, we describe the models used in this work: (i) sequential classifiers without pre-training and (ii) classifiers with pre-training. For the former, we consider long short-term memory networks (LSTM) [143] and convolutional neural networks (CNN) [282]. For the latter, we consider multilingual BERT - mBERT [12] and a Spanish BERT model [285], but we also create a set of specific BERT-based models for Guarani and compare their performance.

Our proposal can be seen as a truly low-resource scenario, where Guarani is not even presented in the mBERT. Note that we only use one GPU and a very small corpus under ∼1M tokens for training.

It should be clear that all models did not take into account any additional information or hand-crafted features and only used raw word inputs, in order to keep a comparable evaluation configuration, although they achieved consistent results.

Neural networks for text classification As baselines, we considered neural networks that process input tweets as a sequence of token vector representations. As mentioned, we consider long short-term memory networks (LSTM) [143] as well as convolutional neural networks (CNN) [282], as implemented in NCRF++ [283]. It should be noted that the former has shown commonly a good performance on sentiment analysis [284]. However, LSTM is usually a standard in many NLP tasks. For the input word embeddings, we used randomly initialized word vectors. We also concatenate a second word embedding, computed through a character CNN (or Bidirectional-LSTM - BiLSTM).

6.3.1 Existing pre-trained language models

As baselines, we fine-tuned contextualized pre-trained language models on the JOTAD training set (see subsection 3.2.3). More particularly, we tested BERT [12] including:

1. *BETO-base*: BETO-base-cased (a Spanish BERT, 12 layers) [285], and
2. *mBERT-base*: multilingual-bert-base-cased (mBERT-base-cased, 12 layers, pre-trained on 104 languages).

It is important to note that in the case of the multilingual model, it is not considered Guarani during pre-training and, that BERT models do not use full tokens, they use a word-piece tokenizer proposed by [288] to generate a vocabulary of the most common sub-word pieces.

Fine-tuning for language modeling We used the weights of the original mBERT and BETO to perform fine-tuning for language modeling. Hoping to obtain the same good results, we follow a similar approach that the adopted for Russian [305] and Portuguese [306], whose models used mBERT as the starting point, too. These models' comparison with a monolingual training with random initialization showed that starting from the pre-trained mBERT, is reduced training time and allows for achieving better performance in various tasks [307]. However, the novelty of our approach lies in doing it in a language that is not in mBERT and with very few resources.

Base models and hyperparameters As for any other language, to train a large language model for Guarani such as BERT, we first need to collect a large set of raw sentences to pre-train the model. More particularly, we rely on the Guarani version of the Wiki-data, an NLP resource widely used by the community. Specifically, we used the

2021-06-02 pages articles[3] dump of the Wikipedia[4] and Wiktionary,[5] both cleaned with `wikiextractor` [308]. We obtained a total of 800K tokens, a much smaller amount than that used for English [12, BERT] or Spanish [285, BETO] with about 3B tokens, and even smaller than other low-resource languages such as Basque [309, BERTeus] with approx. 224M tokens or Galician [307, Bertinho] with about 45M tokens. It should be noted that our proposal is not even comparable to the amount of data used in very resource-poor BERT models for African languages, which used overall 108M tokens. The latter language model, however, is multilingual. While for the training set we used a total of ∼760K tokens, i.e. the first 95% of these wiki articles, the reminder 5% was used for the dev set (about ∼40K tokens). This allowed us to track the *loss* at different training checkpoints to ensure that the training converges.

We then use the Wiki-data to fine-tune the aforementioned models with Guarani. More particularly, we fine-tuned for language modeling the `mBERT-base` and the `BETO-base` models. We used training batches of size 32 considering sequences of 128 tokens. We used the `Hugging Face` library[6] [77] with a single 24GB GPU for training. We trained both models during 1M steps (each took about 5 days to complete). The result is two BERT models fine-tuned for language modeling, namely `mBERT+gn` (mBERT+gn-base-cased) and `BETO+gn` (BETO+gn-base-cased).

6.3.2 Guarani Models

For the pre-training phase, we followed `Bertinho` methodology [307] but, in our case, we trained it on a much smaller corpus, considering a much more restrictive low-resource setup.

Hoping to obtain the same good results as `RoBERTa` [310], we only pre-trained on the masked language objective and ignored the next sentence prediction one.

6.3.2.1 Models

To pre-train our BERT models we followed the procedures, that described below.

BERT tokenizer We defined a sub-word pieces, cased, vocabulary size of 30K for our three models [12, 285, 306, 307], considering words with 2 or more occurrences. In preliminary experiments, we experimented with a vocabulary size of 50K. However, these preliminary tests showed that 30K is a good size due to the morphological correspondence of subwords. Table 6.2 shows an example of a sentence tokenized by mBERT, BETO, and our proposed model.

Pre-training for language modeling As stated above, we have trained three models on Wiki-data, including:

1. *gnBERT-tiny*: BERT-tiny-cased with 2 transformer layers, hidden size set to 256, number of self-attention heads set to 4, and the feed-forward/filter intermediate size set to 768. We decided to use training batches of a size of 192 and considered sequences of 144 tokens.[7] Thus, we trained the model up to 1.5M steps (taking about 10 days to be completed).[8]

[3]Keeping both main texts and the headers.
[4]From `https://dumps.wikimedia.org/gnwiki/`.
[5]From `https://dumps.wikimedia.org/gnwiktionary/`.
[6]`https://github.com/huggingface/transformers`
[7]Note that, a small training batch was used due to our hardware limitations.
[8]We trained the models with more steps, in order to compensate for such a small training batch.

Model	Tokenization
	gn
BETO	Ah ##ata po ##han ##oh ##ár ##ap ##e ha up ##é ##i che ##ró ##ga ##pe .
mBERT	Ah ##ata po ##han ##oh ##ára ##pe ha up ##éi che ##ró ##gap ##e .
Ours	Aha ##ta pohanohára ##pe ha upéi che ##róga ##pe .
	jopara
BETO	Ah ##ata doctor ##pe ha up ##é ##i che ##ró ##ga ##pe .
mBERT	Ah ##ata doctor ##pe ha up ##éi che ##ró ##gap ##e .
Ours	Aha ##ta doctor ##pe ha upéi che ##róga ##pe .

Table 6.2: Tokenization of a sentence in Guarani and Jopara by the original mBERT and BETO and, our model. *Ahata pohanohárape ha upéi cherógape.* (gn). *Ahata doctorpe ha upéi cherógape.* (jopara). *I'm going to the doctor and then to my house.* (en). Thus, BETO's subwords tend to be more shortener and more difficult to interpret, and so are mBERT's subwords. The tokenization of subwords generated by our Guarani BERT seems to take into account inflection and word formation, due to the correspondence with Guarani dictionary words. As it can be seen with `pohanohára` (Doctor of Medicine), `róga` (house), `pe` (to the / to / in), and `upéi` (then / after).

2. *gnBERT-base*: BERT-base-cased with 12 transformer layers, hidden size of 768, the number of self-attention heads on 12, and the feed-forward/filter intermediate size set to 3072. We considered sequences of 128 tokens and used training batches of a size of 96. We trained the model up to 2M steps (took 22 days to complete the training).

3. *gnBERT-large*: BERT-large-cased with 24 transformer layers, hidden size set to 1024, the number of self-attention heads set to 16, and feed-forward/ filter intermediate size of 4096. We used sequences of 128 tokens and considered training batches of 32 of size. So, we trained the model up to 3M steps (completed on 32 days of training).

Although the Guarani Wiki-data is small, expecting the pre-training to converge well and to make sure this happens, we decided to train the models for a long time with more steps compared to [12, 1M steps]. That is, the more complex our models were, the more steps had to be trained.

Our goal is to explore each model's performance according to its architecture's size and complexity in our presented dataset (i.e., JOTAD and JOSA's balanced set).

Hyperparameters We set the hyperparameters and masked language objective following the pre-training configuration and methodology described in [307, p. 17]. We used a learning rate of $1 \times 10 - 4$ with a linear weight decay of 0.01 and, Adam network weight optimizer with $\epsilon = 1 \times 10 - 8$. Given an input sentence, 15% of the tokens was masked randomly, where 80% of them was replaced by the wildcard symbol [MASK], the next 10% was changed to a random word from the input vocabulary, and the remaining 10% are not modified. Again, all our models have been trained using `Hugging Face` transformers library [77] on a single 24GB GPU.

6.4 Experiments and results

Reproducibility The proposed models are available.[9]

[9]Both can be obtained here `https://huggingface.co/mmaguero` and here `https://github.com/mmaguero`. Contact the authors for more details.

We run experiments for the JOTAD dataset, i.e., emotion recognition (JOEMO), humor detection (JOFUN), and offensive language identification (JOFF+) tasks. We evaluated the macro-accuracy to mitigate the impact of (i) the 'e_sad' class in the emotion recognition corpus, and (ii) the distribution of classes in the humor detection and offensive language identification corpora. Additionally, the accuracy, macro-F1 and F1 scores were reported for the JOTAD dataset. In the case of the humor detection and offensive language identification corpora, we calculated the F1 score over the positive classes (i.e., 'fun' and 'off' classes, respectively), while for the emotion recognition corpus, we reported F1 score for the emotion classes ('e_angry', 'e_happy', and 'e_sad').

In all models run, we did a small hyper-parameter search based on the dev set (see details in the Appendix C).

6.4.1 Results and discussion

Table 6.3 shows the results on the emotion recognition task in the JOEMO corpus. Experiments we carried out show that `BETO-base` achieved good results when predicting all classes, but that `gnBERT-tiny` did analogously predicting positive classes, i.e., the emotion classes. In this matter, it is interesting how our 2-layer model obtained better results than the Spanish and multilingual models, each with 12 transformer layers.

Model	**Corpus JOEMO**			
	macro-Acc.	**Acc.**	**macro-F1**	**F1**
Baselines				
^{C}CNN-^{W}BiLSTM	0.4211	0.5015	0.4211	0.4908
^{C}CNN-^{W}CNN	0.5491	0.5428	0.5491	0.5837
^{C}BiLSTM-^{W}CNN	0.5567	0.5809	0.5567	0.5980
^{C}BiLSTM-^{W}BiLSTM	0.5997	0.5841	0.5997	0.5952
BETO$_{base,cased}$	**0.609**	0.654	**0.606**	0.648
mBERT$_{base,cased}$	0.5684	0.6317	0.5644	0.6463
Ours				
BETO+gn$_{base,cased}$	0.5761	0.6349	0.5673	0.6183
mBERT+gn$_{base,cased}$	0.5599	0.6317	0.5637	0.6146
gnBERT$_{tiny,cased}$	0.5766	**0.658**	0.5763	**0.6718**
gnBERT$_{base,cased}$	0.5265	0.6032	0.5409	0.6038
gnBERT$_{large,cased}$	0.25	0.3778	0.1371	0.0

^{C}Encodes character sequence. ^{W}Encodes word sequence.

Table 6.3: Experimental results on emotion recognition task (JOEMO corpus).

With respect to the humor detection task (JOFUN corpus), `gnBERT-tiny` obtained the best results in almost all metrics. The rest of the results are shown in Table 6.4.

Table 6.5 is worth noting that `mBERT+gn` outperformed all the other models on the offensive language identification task (JOFF+ corpus). This suggests that for this task cross-lingual models may improve the performance because `BETO-base` show also concise results.

We tested different combinations with CNNs and BiLSTMs encoders using character representations, which output is first concatenated to a second external word vector (as explained in §6.3), and then fed to the encoder. Among them, the model that used a character-level CNN and a word-level BiLSTM encoder, overall, achieved the best results. In this regard, the results are concise with our findings in chapter 2 and chapter 5.

With respect to the pre-trained language models, they perform better in general than those without pre-training. Hereof, with the exception of the emotion recognition task, our `mBERT+gn` (12 layers) fine-tuned for language modeling, overall obtained robust results. This shows that for a resource-poor linguistic environment, such as Guarani and Jopara,

Model	Corpus JOFUN			
	macro-Acc.	**Acc.**	**macro-F1**	**F1**
		Baselines		
^{C}CNN-^{W}BiLSTM	0.6773	0.7127	0.6773	0.5431
^{C}CNN-^{W}CNN	0.6323	0.7127	0.6323	0.47
^{C}BiLSTM-^{W}CNN	0.6407	0.7046	0.6407	0.4882
^{C}BiLSTM-^{W}BiLSTM	0.6637	0.7615	0.6637	0.5111
BETO$_{base,cased}$	0.6841	0.7263	0.6771	0.5511
mBERT$_{base,cased}$	0.7054	0.6965	0.6708	0.5627
		Ours		
BETO+gn$_{base,cased}$	0.6823	0.7317	0.6786	0.5092
mBERT+gn$_{base,cased}$	0.691	**0.7642**	0.6988	0.5584
gnBERT $_{tiny,cased}$	**0.7208**	0.7344	**0.7**	**0.5984**
gnBERT $_{base,cased}$	0.6963	0.7317	0.6886	0.5677
gnBERT $_{large,cased}$	0.5	0.7458	0.4272	0.0

^{C}Encodes character sequence. ^{W}Encodes word sequence.

Table 6.4: Experimental results on humor detection task (JOFUN corpus).

Model	Corpus JOFF+			
	macro-Acc.	**Acc.**	**macro-F1**	**F1**
		Baselines		
^{C}CNN-^{W}BiLSTM	0.7527	0.8755	0.7527	0.5970
^{C}CNN-^{W}CNN	0.7725	0.841	0.7725	0.5766
^{C}BiLSTM-^{W}CNN	0.7593	0.8479	0.7593	0.5714
^{C}BiLSTM-^{W}BiLSTM	0.7373	0.8594	0.7373	0.5611
BETO$_{base,cased}$	0.8058	0.8871	0.7971	0.6621
mBERT$_{base,cased}$	**0.8269**	0.8548	0.7726	0.6358
		Ours		
BETO+gn$_{base,cased}$	0.7538	0.8871	0.7774	0.6142
mBERT+gn$_{base,cased}$	0.7791	**0.9101**	**0.8127**	**0.6777**
gnBERT $_{tiny,cased}$	0.7626	0.8825	0.7736	0.6165
gnBERT $_{base,cased}$	0.7959	0.8802	0.7859	0.6438
gnBERT $_{large,cased}$	0.5	0.8337	0.4561	0.0

^{C}Encodes character sequence. ^{W}Encodes word sequence.

Table 6.5: Experimental results on offensive language identification task (JOFF+ corpus).

this approach can be very useful.

Following `mBERT+gn`, we find our `gnBERT-tiny` (2 layers) model and the original `BETO-base` release (with 12 layers). With regard to `BETO+gn`, contrary to `mBERT+gn`, we hypothesize that such small data during the language modeling training, harmed the model performance.

As mentioned, `BETO-base` (not pre-trained on Guarani) was performed better than the `BETO+gn`. We think this is partly due to the presence of Spanish tokens in the corpora. This finding is in line with our results reported in chapter 5. It should important to emphasize the cross-lingual abilities of BERT models [289], specifically to deal with this type of situation.

With respect to `gnBERT-large` model, it always predicted only one class (that with more samples). It followed the same behaviour on all the evaluated tasks/corpora, obtaining the worst results. The `gnBERT-large` training presented some instability for the evaluated corpora, which being so small one would expect some degenerative performance during its execution. We believe that this had a negative impact on the fine-tuning process in such a large model. It should be noted that in a shorter fine-tuning process, the original model weights have more influence.

In short, we have trained and evaluated several BERT models for Guarani and Jopara that obtain consistent results against the original mBERT and BETO in our three new

corpora for affect detection. It should highlight that we provided a 2-layer *tiny* BERT, which performed very close to the more complex models and obtained efficient results with a computationally least expensive architecture.

Additional results Since language models using Guarani perform quite well in the task of affect detection, we test it with our corpus proposed in subsection 3.2.2. Our Jopara Sentiment Analysis dataset (JOSA), is a Guarani-dominant polarity classification corpora and is composed of texts from crawled Twitter accounts that usually tweeted in Guarani, which includes both an unbalanced and balanced version.

We run experiments for the balanced version of JOSA dataset and compared it with the three best results reported in Table 5.1 (chapter 5). Remember that we used neural networks models for text classification as in §6.3, and fine-tuned transformer-based models over the base-uncased version. The balanced setup has ∼1.5K tweets with 349 positives, 763 neutrals, and 414 negatives. More particularly, the corpus split is training (50%), development (10%), and test (40%).

Model	Corpus JOSA Balanced
chapter 5*	
^{C}CNN-^{W}BiLSTM	0.57
mBERT$_{base,uncased}$	0.58
BETO$_{base,uncased}$	**0.64**
Ours	
BETO+gn$_{base,cased}$	0.6204
mBERT+gn$_{base,cased}$	**0.6441**
gnBERT $_{tiny,cased}$	0.6041
gnBERT $_{base,cased}$	0.5875
gnBERT $_{large,cased}$	0.3333

^{C}Encodes character sequence. ^{W}Encodes word sequence.

*We reported only two decimal places for the macro-accuracy obtained.

Table 6.6: Experimental results on polarity classification task (JOSA balanced setup). Metric: Macro-accuracy.

It should important to highlight that in this corpus, we only used the macro-accuracy metric, as in chapter 5, mainly for comparison purposes. In Table 6.6 we see that our `mBERT+gn` model barely improved to the `BETO-base` model. The results, which are concise with our findings, indicate that incorporating Guarani into mBERT (or training a specific monolingual-BERT model) is the best option for this type of low-resource environment. It can also be seen again how models with pre-trained, rather than without pre-training, perform better in general.

6.5 Conclusion

In this chapter, we explored Text-based Affect Detection in Jopara, the Guarani-Spanish language code-mixing. We trained and built several BERT-based models in Guarani for this low-resource setting. We followed a truly low-resource approach with only ∼800K tokens for training data on a single 24GB GPU. All models have been evaluated on the JOTAD dataset and, additionally, in the balanced JOSA setting. Specifically, on emotion recognition, humor detection, offensive language identification, and polarity classification. Overall, our mBERT fine-tuned with Guarani - `mBERT+gn` for language modeling, obtains the best results, and even our *tiny* BERT (with only 2 layers) - `gnBERT-tiny` achieved better results than the

original mBERT and BETO on some tasks. In addition, we contributed a Guarani tokenizer that proved to greatly improve the segmentation of Guarani words.

Finally, our experiments have also shown that in a low-resource scenario, a more complex architecture can be hurt the performance in most tasks. We believe this might be due to our limited amount of data and hardware access when training the language model, which forced us to use less complex architectures and transfer-learning approaches, in turn, that resulted in benefits for our experiments. For future work, we aim to extend our models using larger but clean corpora, avoiding noisy datasets. Thus, we hope that the release of the models and corpora will serve to improve the results of NLP applications for Guarani and Jopara.

Acknowledgements

We thank to (i) the Visual Information Processing Group of the University of Granada (especially Javier Mateos) and (ii) the Generalitat Valenciana and the University of Alicante (especially José I. Abreu Salas) through the DGX computing platform (IDIFEDER/2020/003); for giving us access to the GPU hardware necessary to carry out the training of the language models.

Chapter 7

Contributions and Results

This chapter is a summary of the contributions and results obtained in this thesis, which also lists the publications resulting from our work.

7.1 Results

The work presented in this dissertation have contributed to the moving forward the state of the art in the application of sentiment analysis and topic modeling, as well as text mining and related NLP applications, through the formal definition of approaches, their implementation both in study cases and practical tools and methods, while providing linguistic resources for low-resource languages.

7.1.1 Software prototype

`Gastro-miner` is a cloud-based tool that allows end-to-end sentiment analysis of users' reviews and opinions (written in English) about restaurants through travel guidance social media platforms, like TripAdvisor. In its first version, it leverages exclusively on the existing features of TripAdvisor, i.e., entities (restaurants), posts and users. Among other functions (see again 1.9), it allows the (i) data capture and storage, (ii) cleaning and preprocessing of reviews, (iii) sentiment analysis, and (iv) visualization and representation of the analysis and data.

The software tool was developed with a Python stack, namely:

- Scrapy, a web crawling framework for web scraping [311].
- NLTK (Natural Language ToolKit), a toolkit for Natural Language Processing (NLP) provides interfaces to over 50 corpora and lexical resources, along with a suite of text processing libraries (e.g., classification, tokenization, etc.) [312].
- Matplotlib, a 2D plotting library for data visualization [313].
- Django[1], a potent web framework with many plugins that encourages rapid development.

The web environment was based on virtualisation technology, focusing on automation with:

- Vagrant, a tool for building and managing virtual machine environments in an easy-to-use and single workflow [314].

[1] https://www.djangoproject.com/foundation/

- VirtualBox[2], a cross-platform virtualization application.
- Docker stack, for harnessing the benefits of containerization for a focused purpose (i.e., the lightweight packaging and deployment of applications) [315].

For data persistence we used MongoDB, a distributed, document-based, general-purpose database, [316]. We modelled each restaurant as a document. The sentiment analysis stage was based on VADER (Valence Aware Dictionary for sEntiment Reasoning) [221], a rule and lexicon-based sentiment analysis tool. We tokenized user posts into sentences to compute an averaged sentiment. See the full architecture in the Figure 7.1.

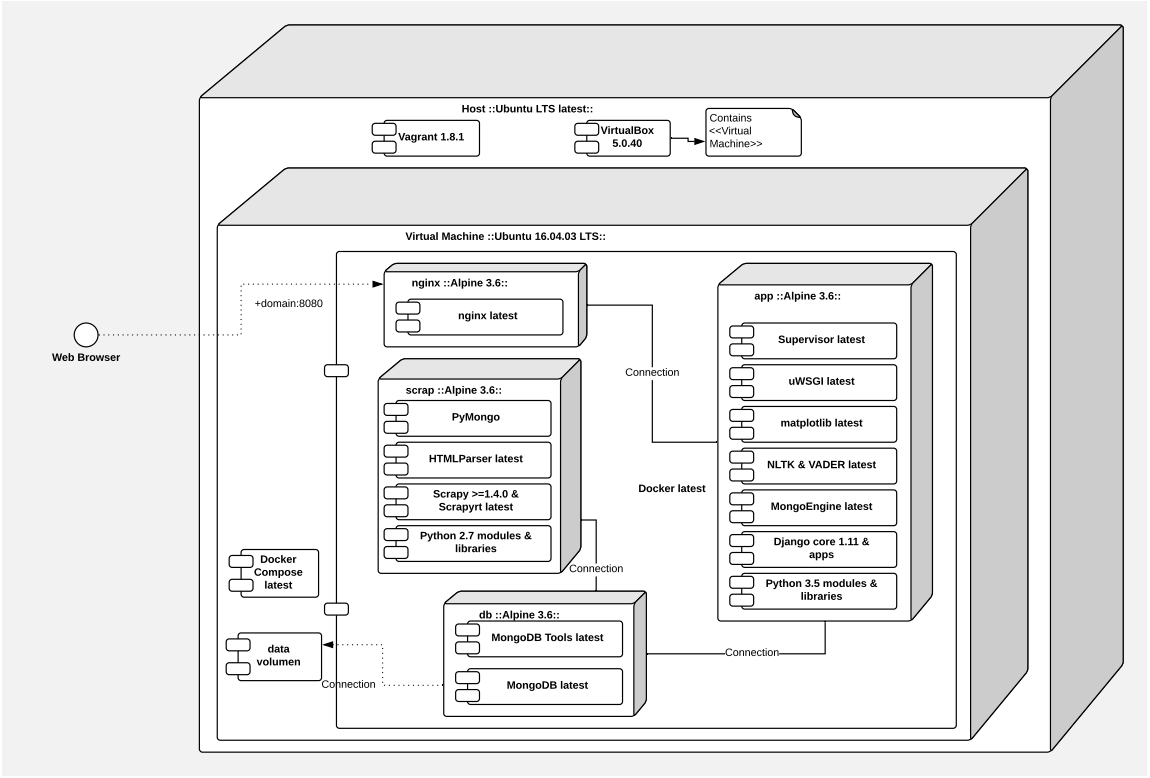


Figure 7.1: General architecture deployed in our `Gastro-miner` tool [91, p. 397, Figure 4].

`Gastro-miner` can be used as a stand-alone application, or it can be customized to other social media, languages, or settings. As we can see in Figure 7.1, each module is straightforwardly interchangeable with any other. We could not test the tool with other social media or languages, dialects, or languages varieties (such as tweets written on low-resourced languages or with code-switching), but still, the methodology, the architecture and the software represent a contribution of this thesis.

The results of this study have been published at *Procedia Computer Science*, was a contribution in *Proceedings of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence* [91]. The section A.1 presents a summary of the tool, followed of the pre-print.

7.1.2 Contributions and resources

As a result of the work carried out in this thesis, and, in addition, of the publications in journals and conferences proceedings resulting from the papers described in Appendix A,

[2]https://www.virtualbox.org/wiki/VirtualBox

different resources and software have been made available to the research community. The main contributions of the thesis are listed below:

1. A cloud-based architecture for a software tool for end-to-end sentiment analysis of social media posts written in English, especially for TripAdvisor restaurant reviews (or any source with similar text-length).

   The complete, cloud-based architecture of the software tool, can be found on Github.[3]

2. Several corpora for low-resource languages:

   (a) An unlabeled Spanish (spoken in Spain) Twitter corpus about the COVID-19 pandemic outbreak, that can be downloaded here.[4]

   (b) The first Guarani-dominant Jopara (Guarani-Spanish code-switching) corpus for sentiment analysis, annotated according to the trinary (positive, negative and neutral) scales. It can be obtained here.[5]

   (c) The first Jopara text-based affect detection dataset can be found here.[6] The dataset includes three multiannotated corpora:

      i. The first Guarani-dominant Jopara corpus for emotion recognition annotated according to four (happy, sad, angry and other) moods.

      ii. The first Guarani-dominant Jopara corpus for humor (humorous) detection.

      iii. The first Guarani-dominant Jopara corpus for offensive and toxic language identification.

3. A small evaluation framework with a small guide (see it in §B.2), which reports the procedure followed by native Spanish-speaking annotators, who were in charge of evaluating a sample of the topics discovered in chapter 4.

4. An annotation mini-guidelines, which describes the process followed by the bilingual annotators (Guarani-Spanish), who manually were annotated the Guarani-dominant Jopara corpora (see it at §B.1).

5. A customized tool composed of several other tools used for language identification. It can be obtained on GitHub.[7]

6. An useful method for discovering topics of Spanish (spoken in Spain) tweets by using linguistic knowledge with generative vs. discriminative routes with the LDA algorithm.

   The topic modeling software on large-data, especially for Twitter, can be found on GitHub.[8]

7. An in-depth review of the works that had leveraged advances on deep learning to tackle the problem of multilingual sentiment analysis in social media, with a high-level perspective of the field, identifying common ideas and problems that have been addressed (valuable for all the public interested in this domain).

8. A detailed and well-structured Twitter dataset creation procedures for code-switching and low-resource languages, covering the limitations encountered during the data gathering process, and the difficulties of collecting it.

[3] `https://github.com/mmaguero/cloud-based-tool-SA`
[4] `https://doi.org/10.7910/DVN/6PPSAZ`
[5] `https://doi.org/10.7910/DVN/GLDX14`
[6] `https://github.com/mmaguero`. Contact the author for more details.
[7] `https://github.com/mmaguero/lang_detection`
[8] `https://github.com/mmaguero/twitter-analysis`

9. A set of pre-trained Guarani language models (BERT-based, a widely used transformer-based model) that can be used to perform a wide range of NLP task in Guarani or Jopara, such sentiment analysis. It can be obtained at Hugging Face Models Hub:[9]

   (a) Three monolingual BERT [12] models for the Guarani language, trained with the data from Wikipedia in Guarani.

   (b) Two BERT language models fine-tuned with data from Wikipedia in Guarani, namely Spanish BERT [285, BETO] and mBERT (multilingual BERT) [12].

10. A Guarani BERT tokenizer, which can also be found on the Hugging Face Models Hub.

Note that from the above corpora only the tweets-ids with tags/labels are given, for a research version (i.e., the version of the corpus with tweets), contact the author of this dissertation.[10]

7.2 List of publications

A list of the publications of this thesis is given below. To see the preprints and their summary, please go to Appendix A.

Agüero-Torales, M. M., López-Herrera, A. G. & Cobo, M. J. (2018). GASTRO-MINER - Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor: Caso de Estudio sobre Restaurantes de la Provincia de Granada. In LIBRO DE RESÚMENES - I Jornadas Científicas en CIENCIA DE DATOS, (pp. 34). Universidad Comunera. First place in the 'i-Data' contest. doi: http://dx.doi.org/10.13140/RG.2.2.33136.71688.[11]

Agüero-Torales, M. M., Cobo, M. J., Herrera-Viedma, E., & López-Herrera, A. G. (2019). A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor. Procedia Computer Science, 162, 392-399. doi: https://doi.org/10.1016/j.procs.2019.12.002.

Agüero-Torales, M. M., Vilares, D., & López-Herrera, A. G. (2021). Discovering topics in Twitter about the COVID-19 outbreak in Spain. Procesamiento del Lenguaje Natural, 66, 177-190. URL:http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6333.

Agüero-Torales, M. M., Abreu Salas, J. I., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 107373. doi: https://doi.org/10.1016/j.asoc.2021.107373.

Agüero-Torales, M., Vilares, D., & López-Herrera, A. (2021, June). On the logistical difficulties and findings of Jopara Sentiment Analysis. In Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching (pp. 95-102). Association for Computational Linguistics. doi: http://dx.doi.org/10.18653/v1/2021.calcs-1.12.

[9] https://huggingface.co/mmaguero. Contact the author for more details.

[10] Preferably to maguero(at)correo(dot)ugr(dot)es; marvin(hyphen)aguero(at)hotmail(dot)com.

[11] https://www.ucom.edu.py/wp-content/uploads/2019/08/Jornadas-Cientificas-en-Ciencia-de-Datos-UCOM.pdf#page=34

Chapter 8

Conclusion

This chapter presents a summary of the conclusions obtained. In addition, it points out some open lines of work derived from the results obtained.

8.1 Concluding Remarks

In this thesis we have addressed several issues that pursue a common goal: to develop methods, approaches and resources to deal with multilingualism and resource-poor languages on social media, providing insights, specially into the area of topic modeling and sentiment analysis, a better understanding of this problem as well as a holistic analysis and view of it. More particularly, two main objectives have been pursued in this thesis:

1. Study the different existing machine-learning approaches, that aim the language-independence, especially the most current ones based on neural networks, with respect to multilingual opinions written in social media, even if was written with code-switching.

2. Develop new linguistic resources for text analysis in low-resource languages and dialects, especially the written in social media.

3. Build machine learning models for NLP of low-resource languages and dialects, especially in monolingual, multilingual, and code-switching settings.

It is necessary to remark our research highlights for the chapter 2: (i) a review of Deep Learning applications to address Multilingual Sentiment Analysis, (ii) a fast-growing and clear interest in deep-learning based models for (agnostic) multilingual sentiment analysis, with 24 papers from 2017 to 2020, (iii) papers covering 23 different languages and 11 social media datasets or corpora, (iv) mixed performance of models applied by the research community, but word embeddings and CNN or LSTM were trending options at the time (with a noticeable migration to transformer-based models), (v) embedding $\rangle$ feature extractor $\rangle$ classifier, was the predominant architecture, except concerning aspect sentiment analysis (as this usually requires a more complex architecture).

This literature review focused on and highlighted three main aspects of each work we revised: (i) the underlying hypothesis relative to how they handled multilingualism, (ii) the design decisions to implement the hypothesis, and (iii) the experimental results that evaluated the hypothesis. Consequently, in this work, we not only discussed our main findings but also highlighted some unexplored topics that may hint at interesting directions for further research. Thereby, this review work regarding relevant issues, like the increase of demand of solving more difficult sentiment analysis tasks, while deep learning algorithms with more elaborated architectures are required to solve them. More particularly, an active

challenge in sentiment analysis, and many NLP tasks, is precisely, the unfairness about the development of resources for different languages, especially for the poor resource ones. Lastly, our review may be of interest to those researchers who intend to employ a more complex Multilingual Sentiment Analysis agnostic approach.

We have seen in chapter 4, that although qualitative evaluation is a good complement to analyze the results of a system, it is not an objective tool to measure the quality of the contributions. Therefore, we also performed a small human evaluation framework to quantitatively estimate the quality of the extracted topics, since it was difficult to find works that made topic modeling in Spanish about the COVID-19 (and especially the one spoken in Spain). This evaluation allowed us to measure the quality of the terms extracted by the system. In fact, due to the scarcity of these works, our work proposed an easy-to-read approach, which detailed it in a very clear way following a strict methodology. We conclude that tuning a well-known algorithm, such LDA, is feasible way to perform robust analysis when discovering topics in short texts, such as tweets. Besides, we provided a methodology to collect domain-specific geolocalized public dataset (see subsection 3.2.1).

The work presented in subsection 3.2.2 and chapter 5, was a well-documented and easy-to-read exploration of sentiment analysis in a low-resource language such Guarani. Previous approaches to sentiment analysis in Guarani [40], only covered Jopara, resulting in fact, in datasets dominated by Spanish. Our work, in contrast, ensured that the dataset was Guarani and Jopara dominant, resulting in the first Guarani-dominant Jopara sentiment analysis dataset. In addition, our work provided an interesting and very detailed description of the limitations encountered during the data collection process, the difficulties of collecting Guarani-dominant and Jopara-dominant tweets, and even of performing machine learning approaches in this very particular environment, and how these were overcome. This may be useful for other researchers tackling the construction of datasets of this type.

In chapter 6 we pre-trained a set of BERT models for an indigenous and low-resource language such Guarani. Note that existing multilingual transformer-based models do not cover Guarani (e.g., mBERT or XLM-R, see subsection 1.2.2 for more details), for why our proposed models, resulting in fact, in the first approach for this type of language. On the other hand, the corpus created in subsection 3.2.3 is the first Guarani-dominant Jopara text-based affect detection dataset. Note that the dataset was created in a multi-annotation task, which is not very common, let alone in low-resource and code-switching scenarios. We believe this contribution should help to the democratization of these languages, Guarani and Jopara. At the same time, this contribution is an interesting and detailed approach for how to overcome hardware limitations as well as data-amount issues for training an efficient BERT model for a low-resource language.

Finally, it should be noted that opinions are important in public opinion studies, negative campaigning or political polarization, and topic discovery, to name a few contexts, providing ample opportunities for innovative research in communities with low-resource languages speakers, such as Spanish spoken in Spain, Guarani or Jopara, and indeed the democratization of such languages. We believe that our work approaches solutions in a concise, conceptual, and practical manner to the above points, expanding the state-of-the-art.

Conclusiones

En esta tesis hemos abordado varias cuestiones que persiguen un objetivo común: desarrollar métodos, enfoques y recursos para tratar el multilingüismo y las lenguas con escasez de recursos en los medios sociales, proporcionando, especialmente en el ámbito del análisis sentimientos y el modelado de temas, una mejor comprensión de este problema, así como un análisis y una visión global del mismo. Más concretamente, en esta tesis se han perseguido dos objetivos principales:

1. Estudiar los distintos enfoques existentes de aprendizaje automático, que tienen como objetivo la independencia lingüística, en especial los más actuales basados en redes neuronales, respecto a comentarios multilingües y de *code-switching* procedentes de medios sociales.

2. Desarrollar nuevos recursos lingüísticos para el análisis de comentarios en idiomas y dialectos *low-resource*, especialmente los escritos en medios sociales.

3. Construir modelos de aprendizaje automático para el PLN de idiomas y dialectos *low-resource*, especialmente en entornos monolingües, multilingües y de cambio de código.

Es necesario señalar los aspectos más destacados de nuestra investigación para el capítulo 2: (i) una revisión de las aplicaciones de *Deep Learning* para abordar el Análisis de Sentimiento Multilingüe, (ii) un rápido crecimiento y claro interés en los modelos basados en aprendizaje profundo para el análisis de sentimiento multilingüe (de una manera agnóstica), con 24 artículos desde 2017 hasta 2020, (iii) artículos que cubren 23 idiomas diferentes y 11 conjuntos de datos o corpus de medios sociales, (iv) rendimiento mixto de los modelos aplicados por la comunidad investigadora, pero las incrustaciones de palabras (también conocidas como *embeddings*) y las redes del tipo CNN o LSTM eran las opciones de tendencia en ese momento (con una notable migración a los modelos basados en *transformers*), (v) *embeddings* $\rangle$ extractor de características $\rangle$ clasificador, era la arquitectura predominante, excepto en lo que respecta al análisis de sentimiento basado en aspectos (ya que esto generalmente requiere una arquitectura más compleja).

Esta revisión de la literatura se centró y destacó tres aspectos principales de cada trabajo que revisamos: (i) la hipótesis subyacente relativa a cómo trataban el multilingüismo, (ii) las decisiones de diseño para implementar la hipótesis, y (iii) los resultados experimentales que evaluaban la hipótesis. En consecuencia, en este trabajo no sólo discutimos nuestros principales hallazgos, sino que también destacamos algunos temas inexplorados que pueden sugerir direcciones interesantes para futuras investigaciones. Por lo tanto, este trabajo de revisión se refiere a cuestiones relevantes, como el aumento de la demanda de resolución de tareas de análisis de sentimientos más difíciles, mientras que se requieren algoritmos de aprendizaje profundo con arquitecturas más elaboradas para resolverlas. Más concretamente, un reto activo en el análisis de sentimientos, y en muchas tareas de PNL, es precisamente, la falta de equidad en el desarrollo de recursos para diferentes idiomas, especialmente para los de escasos recursos. Por último, nuestra revisión podría ser de interés para aquellas investigaciones que pretendan emplear un enfoque agnóstico de Análisis de Sentimiento Multilingüe más complejo.

Hemos visto en el capítulo 4, que aunque la evaluación cualitativa es un buen complemento para analizar los resultados de un sistema, no es una herramienta objetiva para medir la calidad de las aportaciones. Por lo tanto, también realizamos un pequeño *framework* de evaluación humana para estimar de forma cuantitativa la calidad de los temas extraídos, ya que era difícil encontrar trabajos que hicieran modelado de temas en español sobre COVID-19 (y especialmente la variante del español que se habla en España). Esta evaluación nos permitió medir la calidad de los términos extraídos por el sistema. De hecho,

debido a la escasez de estos trabajos, nuestro trabajo propuso un enfoque de fácil lectura, que lo detallaba de forma muy clara siguiendo una metodología estricta. Concluimos que afinar un algoritmo bien conocido, como el LDA, es una forma viable de realizar un análisis robusto a la hora de descubrir temas en textos cortos, como los tweets. Además, en la sección 3.2.1 proporcionamos una metodología para recopilar un conjunto de datos públicos geolocalizados específicos del dominio.

El trabajo presentado en la sección 3.2.2 y el capítulo 5 fue una exploración bien documentada y fácil de leer del análisis de sentimientos en una lengua de bajos recursos como el guaraní. Las aproximaciones anteriores al análisis de sentimientos en guaraní [40], sólo cubrían el jopará, resultando de hecho, en conjuntos de datos dominados por el español. Nuestro trabajo, por el contrario, garantizó que el conjunto de datos fuera dominante en guaraní y en jopará, lo que dio como resultado el primer conjunto de datos de análisis de sentimientos en jopará dominante en guaraní. Además, nuestro trabajo proporcionó una descripción interesante y muy detallada de las limitaciones encontradas durante el proceso de recopilación de datos, las dificultades de recoger tuits con predominio del guaraní y el jopará, e incluso de realizar enfoques de aprendizaje automático en este entorno tan particular, y cómo se superaron. Esto podría ser útil para otros investigadores que aborden la construcción de *datasets* de esta tipología.

En el capítulo 6 preentrenamos un conjunto de modelos BERT para una lengua autóctona y de bajos recursos como el guaraní. Cabe destacar que los modelos multilingües existentes basados en *transformers* no cubren el guaraní (por ejemplo, mBERT o XLM-R, véase el apartado 1.2.2 para más detalles), por lo que nuestros modelos propuestos, resultan de hecho, en la primera aproximación para este tipo de lengua. Por otro lado, el corpus creado en la sección 3.2.3 es el primer conjunto de datos de *affect detection* basados en texto guaraní y jopará. Cabe destacar que el conjunto de datos fue creado en una tarea de anotación múltiple, lo cual no es muy común, y menos aún en escenarios de bajos recursos y de *code-switching*. Creemos que esta contribución debería ayudar a la democratización de estas lenguas, es decir, al guaraní y al jopará. Al mismo tiempo, esta contribución es un enfoque interesante y detallado de cómo superar las limitaciones de hardware, así como los problemas de escasez de datos para el entrenamiento de un modelo BERT para una lengua *low-resource*.

Por último, cabe señalar que las opiniones son importantes en los estudios de opinión pública, en las campañas negativas o en la polarización política, y en el descubrimiento de temas, por nombrar algunos contextos, lo que ofrece amplias oportunidades para la investigación innovadora en comunidades con hablantes de lenguas de bajos recursos, como el español hablado en España, el guaraní o el jopará, y de hecho, la democratización de dichas lenguas. Creemos que nuestro trabajo aborda soluciones de manera concisa, conceptual y práctica a los puntos anteriores, ampliando el estado del arte.

8.2 Future Work

The resources, results, methodologies, datasets, and other materials made available in this thesis (see chapter 7) can encourage researchers to create new linguistic resources for different resource-poor languages, especially for the 'truly low-resource' languages, and in particular indigenous languages, as well as code-switching languages. Linguistic resources are fundamental to performing all NLP tasks. Therefore, We would like to continue contributing future resources for the Guarani language and Jopara, including the incorporation of domain-specific resources, such as those for the health, political, or business sectors, to name a few. On the other hand, we also aim to study the viability of the Aspect-Based Sentiment Analysis (ABSA) to these multilingual and low-resource contexts.

Even though in this thesis we have proposed useful systems, approaches, methodologies, and resources for solving sentiment analysis and topic modeling problems in low-resource languages settings, there are still some challenges to overcome. For example, in chapter 4, we tuned a LDA-based system for topic discovery in opinions expressed in Spanish. As future work, we plan to study the applicability of the above with truly low-resource languages such as Guarani or Jopara. In addition, we intend to experiment with more topic model approaches, like the transformed-based, or even explore hierarchical topic models in order to identify subtopics automatically [317].

On the other hand, we desire to develop a more efficient monolingual BERT model for Guarani than the proposed in chapter 6. Our approach showing that it is feasible to build robust monolingual BERT models even for low-resource languages. This finding is in line with the recent trend [307, 309], achieving better performance than the official multilingual BERT (mBERT). But Guarani is a truly resource-poor language with a very small Wikipedia size, a fact which did not make it possible to achieve better results such as the aforementioned above. However, our built BERT models have proved slightly better than the original mBERT or Spanish BERT (BETO). Our experiments show that transfer learning approaches, i.e., transfer languages for cross-lingual learning, are useful for low-resource languages, especially using a related language to the target language, outperforming many times the results in our studied tasks (see chapter 5 and chapter 6). The truth is that we need more training data to improve the monolingual BERT model as well as the transfer learning models, which does not mean trying to tackle other less data-hungry approaches, like Fine-tuned LAnguage Net (FLAN) [318], which involves fine-tuning a model not to solve a specific task, but to make it more amenable to solving NLP tasks in general. Precisely, with the same idea, we can also propose in the future a model applicable to all majority of languages applying zero-shot learning approaches [319, 320].

About model exploration, we think it may be worthwhile to look beyond the standard SOTA Transformer model. For example, UTs (Universal Transformers) [321], that generalizes Transformers with recurrent connections. Moreover, in all our work, all labels are predicted in an atomic way (i.e., single-task learning), so, we propose to explore multi-task learning (MTL) approaches over our created corpora, which have already proven to be beneficial in affect detection [322] and low-resource languages settings [323, 324, 325].

Finally, as future work (as mentioned in subsection 7.1.1), we plan to update our cloud-based tool proposed in section A.1 [91, 92], incorporating the data science life-cycle workflow (see Figure 1.10 in section 1.6), which was used throughout this thesis, and which is adapted to textual content written in low-resource languages on social media platforms. Thus, our tool could address both sentiment analysis and topic modeling in low-resource languages and multilingual environments (especially for code-switching), so achieving an efficient and robust end-to-end tool for social media opinion analysis.

Bibliography

[1] Simon Kemp. Half a billion users joined social in the last year (and other facts), Jul 2021.

[2] Ethnologue. What are the top 200 most spoken languages. *Ethnologue: Languages of the World. Twenty-fourth edition*, 2021.

[3] Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach, global edition 4th. *Foundations*, 19:23, 2021.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[5] Xin-She Yang. 2 - mathematical foundations. In Xin-She Yang, editor, *Introduction to Algorithms for Data Mining and Machine Learning*, pages 19–43. Academic Press, 2019.

[6] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal, editors. *Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition)*. Morgan Kaufmann, fourth edition, 2017.

[7] Steven L Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, 1994.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] FBBVA and David Vilares. Beca Leonardo a Investigadores y Creadores Culturales - David Vilares Calvo, Tecnologías de la Información y la Comunicación, 2020. *FBBVA*, 2020.

[10] FBBVA and David Vilares. David Vilares Calvo - Premio de Investigación Sociedad Científica Informática de España-Fundación BBVA, Jóvenes Investigadores Informáticos, 2018. *FBBVA*, 2018.

[11] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[14] Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. MetaXL: Meta representation transformation for low-resource cross-lingual learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online, June 2021. Association for Computational Linguistics.

[15] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[16] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017.

[17] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[18] Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online, June 2021. Association for Computational Linguistics.

[19] Jeffrey F. Cohn, Fernando De la Torre, Nadia Bianchi-Berthouze, Andrea Kleinsmith, Chi-Chun Lee, Jangwon Kim, Carlos Busso Angeliki Metallinou, Sungbok Lee, Shrikanth S. Narayanan, Carlo Strapparava, Rada Mihalcea, Jennifer Healey, Christian Mühl, Dirk Heylen, Anton Nijholt, Ryan S. J. D. Baker, Jaclyn Ocumpaugh, Ginevra Castellano, Hatice Gunes, Christopher Peters, and Björn Schuller. *Affect Detection - Section 2*, chapter 10–17. Oxford Library of Psychology, 2015. The Oxford Handbook of Affective Computing.

[20] Jesus Serrano-Guerrero, Antonio Gabriel Lopez-Herrera, Pablo Jimenez, José Angel Olivas, and Enrique Herrera-Viedma. Fuzzy methodology for recommendation based on sentiment analysis and content tools. In *SoMeT*, pages 285–298, 2018.

[21] Hamed Yaghoobian, Hamid R. Arabnia, and Khaled Rasheed. Sarcasm detection: A comparative study, 2021.

[22] Paul Ekman. Facial expressions of emotion: New findings, new questions, 1992.

[23] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

[24] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. psychology press, 2001.

[25] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[26] Flor Miriam Plaza del Arco, Salud María Jiménez-Zafra, Arturo Montejo-Ráez, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Overview of the emoevales task on emotion detection for spanish at iberlef 2021. *Procesamiento del Lenguaje Natural*, 67(0):155–161, 2021.

[27] Abeer AlDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021.

[28] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

[29] Luis Chiruzzo, Santiago Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In *IberLEF@ SEPLN*, pages 132–144, 2019.

[30] Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[31] Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosa, J. A. Meaney, and Rada Mihalcea. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67(0):257–268, 2021.

[32] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[33] Flor Miriam Plaza del Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021.

[34] Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, 2018.

[35] Juan Bernabé Moreno et al. *New methods for knowledge discovery in geo-localized social media networks*. PhD thesis, Universidad de Granada, 2016.

[36] Valentin Barriere and Alexandra Balahur. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 266–271, 2020.

[37] Fahim Djatmiko, Ridi Ferdiana, and Muhammad Faris. A review of sentiment analysis for non-english language. In *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, pages 448–451. IEEE, 2019.

[38] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069, 2019.

[39] Carol Myers-Scotton. Code-switching. *The Handbook of Sociolinguistics*, pages 217–237, 2017.

[40] Adolfo A Ríos, Pedro J Amarilla, and Gustavo A Giménez Lugo. Sentiment categorization on a creole language with lexicon-based and machine learning techniques. In *2014 Brazilian Conference on Intelligent Systems*, pages 37–43. IEEE, 2014.

[41] Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi Gonzàlez. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. *Journal of medical internet research*, 22(7):e17758, 2020.

[42] David Vilares, Mike Thelwall, and Miguel A Alonso. The megaphone of the people? spanish sentistrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6):799–813, 2015.

[43] Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348, 2021.

[44] F Steiner-Correa, María I Viedma-del Jesus, and Antonio Gabriel Lopez-Herrera. A survey of multilingual human-tagged short message datasets for sentiment analysis tasks. *Soft Computing*, 22(24):8227–8242, 2018.

[45] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[46] Janyce Wiebe, Rebecca Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 246–253, 1999.

[47] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th annual meeting on ACL*, pages 417–424. ACL, 2002.

[48] David Vilares. *Compositional language processing for multilingual sentiment analysis*. PhD thesis, Universidade da Coruña, 2017.

[49] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[50] Federico Pascual. Topic modeling: An introduction, 2019.

[51] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[52] Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[53] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[54] Christopher E Moody. Mixing dirichlet topic models and word embeddings to make lda2vec, 2016. Submitted to CoNLL 2016.

[55] Nicole Peinelt, Dong Nguyen, and Maria Liakata. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, 2020.

[56] Dimo Angelov. Top2vec: Distributed representations of topics, 2020.

[57] Maarten Grootendorst and Nils Reimers. Maartengr/bertopic: v0.9.3 - quickfix, October 2021.

[58] W John Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer, 2004.

[59] Michael D Gordin. *Scientific babel*. University of Chicago Press, 2015.

[60] Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online, June 2021. Association for Computational Linguistics.

[61] John R Pierce and John B Carroll. Language and machines: Computers in translation and linguistics. Technical report, National Academy of Sciences/National Research Council, 1966.

[62] Margalit Fox. A changed noam chomsky simplifies. *The New York Times*, 1998.

[63] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[64] Keith D. Foote. A brief history of natural language processing (nlp), 2019.

[65] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

[66] Oliver Lemon and Olivier Pietquin, editors. *Conversational Interfaces*, pages 1–4. Springer New York, New York, NY, 2012.

[67] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[68] Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. Character-aware neural language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[69] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[70] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, number 9, pages 1045–1048. International Speech Communication Association, 2010.

[71] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278, 2013.

[72] Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018.

[73] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[75] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[76] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online, October 2020. Association for Computational Linguistics.

[77] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[78] Kristin A Cook and James J Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.

[79] John Lehrberger and Laurent Bourbeau. *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation.* John Benjamins, 1988.

[80] Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 99–103, 2018.

[81] Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

[82] Alexander J. Ratner, Stephen H. Bach, Henry R. Ehrenberg, and Chris Ré. Snorkel: Fast training set generation for information extraction. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1683–1686, New York, NY, USA, 2017. Association for Computing Machinery.

[83] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018.

[84] Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors. *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, Online, November 2020. Association for Computational Linguistics.

[85] Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Luis Gascó, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 13–20, 2021.

[86] Touseef Iqbal and Shaima Qureshi. The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, 2020.

[87] Horacio Saggion. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137, 2017.

[88] Daniel Ferrés, Montserrat Marimon, Horacio Saggion, and Ahmed AbuRa'ed. Yats: Yet another text simplifier. In Elisabeth Métais, Farid Meziane, Mohamad Saraee, Vijayan Sugumaran, and Sunil Vadera, editors, *Natural Language Processing and Information Systems*, pages 335–342, Cham, 2016. Springer International Publishing.

[89] Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–208, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[90] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges, 2020.

[91] M.M. Agüero-Torales, M.J. Cobo, E. Herrera-Viedma, and A.G. López-Herrera. A cloud-based tool for sentiment analysis in reviews about restaurants on tripadvisor. *Procedia Computer Science*, 162:392–399, 2019. 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence.

[92] Marvin M. Aguero-Torales, A. G. López-Herrera, and Manuel J. Cobo. Gastro-miner: Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor. Caso de Estudio sobre Restaurantes de la Provincia de Granada. In *LIBRO DE RESÚMENES - I Jornadas Científicas en CIENCIA DE DATOS*, page 34, Asunción, Paraguay, October 2018. Universidad Comunera. Abstract (Poster).

[93] Chris Richardson. *Microservices patterns*. Manning Publications, November 2018.

[94] Marvin M. Agüero-Torales, José I. Abreu Salas, and Antonio G. López-Herrera. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, 107:107373, 2021.

[95] Marvin M. Agüero-Torales, David Vilares, and Antonio G. López-Herrera. Discovering topics in twitter about the covid-19 outbreak in spain. *Procesamiento del Lenguaje Natural*, 66(0):177–190, 2021.

[96] Marvin Agüero-Torales, David Vilares, and Antonio López-Herrera. On the logistical difficulties and findings of jopara sentiment analysis. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 95–102, Online, June 2021. Association for Computational Linguistics.

[97] Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4):499–527, Dec 2017.

[98] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[99] Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, pages 90–94. Association for Computational Linguistics, 2012.

[100] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[101] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[102] Cicero Dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.

[103] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[104] Duyu Tang, Bing Qin, and Ting Liu. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303, 2015.

[105] Lina Maria Rojas-Barahona. Deep learning for sentiment analysis. *Language and Linguistics Compass*, 10(12):701–719, 2016.

[106] Prerana Singhal and Pushpak Bhattacharyya. Sentiment analysis and deep learning: a survey. *Center for Indian Language Technology, Indian Institute of Technology, Bombay*, 2016.

[107] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A Rehman. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, 8(6):424, 2017.

[108] Lei Johnny Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8, 2018.

[109] Olivier Habimana, Yuhua Li, Ruixuan Li, Xiwu Gu, and Ge Yu. Sentiment analysis using deep learning approaches: an overview. *Science China Information Sciences*, 63(1):1–36, 2019.

[110] Ramesh Wadawadagi and Veerappa Pagi. Sentiment analysis with deep neural networks: comparative study and performance assessment. *Artificial Intelligence Review*, 2020.

[111] Hitesh Nankani, Hritwik Dutta, Harsh Shrivastava, PVNS Rama Krishna, Debanjan Mahata, and Rajiv Ratn Shah. Multilingual sentiment analysis. In *Deep Learning-Based Approaches for Sentiment Analysis*, pages 193–236. Springer, 2020.

[112] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July 2002.

[113] Ronen Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, 2013.

[114] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21, 2013.

[115] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*, 2013.

[116] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[117] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[118] Jianfei Yu and Jing Jiang. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246, Austin, Texas, November 2016. Association for Computational Linguistics.

[119] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[120] Binxuan Huang, Yanglan Ou, and Kathleen M Carley. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer, 2018.

[121] Yukun Ma, Haiyun Peng, and Erik Cambria. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[122] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, 2016.

[123] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

[124] Guozheng Rao, Weihang Huang, Zhiyong Feng, and Qiong Cong. Lstm with sentence representations for document-level sentiment classification. *Neurocomputing*, 308:49 – 57, 2018.

[125] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org, 2015.

[126] Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243, 2017.

[127] Basant Agarwal, Richi Nayak, Namita Mittal, and Srikanta Patnaik. *Deep learning-based approaches for sentiment analysis*. Springer, 2020.

[128] Jonatas Wehrmann, Willian Becker, Henry E. L. Cagnini, and Rodrigo C. Barros. A character-based convolutional neural network for language-agnostic twitter sentiment analysis. *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2384–2391, 2017.

[129] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[130] Guangyou Zhou, Zhao Zeng, Jimmy Xiangji Huang, and Tingting He. Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 245–254. ACM, 2016.

[131] Prerana Singhal and Pushpak Bhattacharyya. Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3053–3062, 2016.

[132] Zhongqing Wang, Yue Zhang, Sophia Lee, Shoushan Li, and Guodong Zhou. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1634, 2016.

[133] Souvick Ghosh, Satanu Ghosh, and Dipankar Das. Sentiment identification in code-mixed social media text, 2017.

[134] Gazi Imtiyaz Ahmad, Jimmy Singla, and Nikita Nikita. Review on sentiment analysis of indian languages with a special focus on code mixed indian languages. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pages 352–356. IEEE, 2019.

[135] Erik Tromp. *Multilingual sentiment analysis on social media*. Lap Lambert Academic Publ, 2012.

[136] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual sentiment and subjectivity analysis. *Multilingual natural language processing*, 6:1–19, 2011.

[137] Iñaki San Vicente Roncal. *Multilingual sentiment analysis in social media*. PhD thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2019.

[138] Malak A. Abdullah. *Deep Learning for Sentiment and Emotion Detection in Multilingual Contexts*. PhD thesis, The University of North Carolina at Charlotte, 2018. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Última actualización - 2019-10-18.

[139] Betül Ay Karakuş, Muhammed Talo, İbrahim Rıza Hallaç, and Galip Aydin. Evaluating deep learning models for sentiment classification. *Concurrency and Computation: Practice and Experience*, 30(21):e4783, 2018.

[140] Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th international conference on world wide web*, pages 1045–1052. International World Wide Web Conferences Steering Committee, 2017.

[141] Willian Becker, Jonatas Wehrmann, Henry EL Cagnini, and Rodrigo C Barros. An efficient deep neural architecture for multilingual sentiment analysis in twitter. In *The Thirtieth International Flairs Conference*, 2017.

[142] Igor Mozetič, Miha Grčar, and Jasmina Smailović. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

[143] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[144] Shiwei Zhang, Xiuzhen Zhang, and Jeffrey Chan. Language-independent twitter classification using character-based convolutional networks. In *International Conference on Advanced Data Mining and Applications*, pages 413–425. Springer, 2017.

[145] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *CSCW*, 2015.

[146] Shiwei Zhang, Xiuzhen Zhang, and Jeffrey Chan. A word-character convolutional neural network for language-agnostic twitter sentiment analysis. In *Proceedings of the 22nd Australasian Document Computing Symposium*, page 12. ACM, 2017.

[147] Lisa Medrouk and Anna Pappa. Deep learning model for sentiment analysis in multilingual corpus. In *International Conference on Neural Information Processing*, pages 205–212. Springer, 2017.

[148] Lisa Medrouk and Anna Pappa. Do deep networks really need complex modules for multilingual sentiment polarity detection and domain classification? In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[149] Guangfeng Liu, Xianying Huang, Xiaoyang Liu, and Anzhi Yang. A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network. *The Computer Journal*, 2019.

[150] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30, 2016.

[151] Xin Dong and Gerard De Melo. Cross-lingual propagation for deep sentiment analysis. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[152] Konstantinos Stavridis, Georgia Koloniari, and Euclid Keramopoulos. Deriving word embeddings using multilingual transfer learning for opinion mining. In *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*, pages 1–6. IEEE, 2018.

[153] Weichao Wang, Shi Feng, Wei Gao, Daling Wang, and Yifei Zhang. Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, 2018.

[154] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 557–570, 2018.

[155] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, 2015.

[156] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, 2013.

[157] Jianghong Shen, Xiaodong Liao, and Shuai Lei. Cross-lingual sentiment analysis via aae and bigru. In *2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 237–241. IEEE, 2020.

[158] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.

[159] Mohammed Jabreel, Najlaa Maaroof, Aïda Valls, and Antonio Moreno. Unisent: Universal sentiment analysis system for low-resource languages. In *CCIA*, pages 387–396, 2019.

[160] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

[161] Kamil Kanclerz, Piotr Miłkowski, and Jan Kocoń. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Computer Science*, 176:128–137, 2020.

[162] Rouzbeh Ghasemi, Seyed Arad Ashrafi Asli, and Saeedeh Momtazi. Deep persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal of Information Science*, page 0165551520962781, 2020.

[163] Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. Sentiment analysis of code-mixed languages leveraging resource rich languages. In *19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2018)*, 2018.

[164] Arouna Konate and Ruiying Du. Sentiment analysis of code-mixed bambara-french social media text using deep learning techniques. *Wuhan University Journal of Natural Sciences*, 23(3):237–243, 2018.

[165] Madan Gopal Jhanwar and Arpita Das. An ensemble model for sentiment analysis of hindi-english code-mixed data, 2018.

[166] Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, 2016.

[167] K Shalini, HB Barathi Ganesh, M Anand Kumar, and KP Soman. Sentiment analysis for code-mixed indian social media text with distributed representation. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1126–1131. IEEE, 2018.

[168] Braja Gopal Patra, Dipankar Das, and Amitava Das. Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task @ICON-2017, 2018.

[169] Yash Kumar Lal, Vaibhav Kumar, Mrinal Dhar, Manish Shrivastava, and Philipp Koehn. De-mixing sentiment from code-mixed text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 371–377, 2019.

[170] Siddhartha Mukherjee. Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4. IEEE, 2019.

[171] Anupam Jamatia, Steve Swamy, Björn Gambäck, Amitava Das, and Swapam Debbarma. Deep learning based sentiment analysis in a code-mixed english-hindi and english-bengali social media corpus. *International journal on artificial intelligence tools*, 29(5), 2020.

[172] Ramesh Chundi, Vishwanath R Hulipalled, and JB Simha. Saekcs: Sentiment analysis for english–kannada code switchtext using deep learning techniques. In *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pages 327–331. IEEE, 2020.

[173] Yonghua Zhu, Xun Gao, Weilin Zhang, Shenkai Liu, and Yuanyuan Zhang. A bi-directional lstm-cnn model with attention for aspect-level text classification. *Future Internet*, 10(12):116, 2018.

[174] Benjamin Heinzerling and Michael Strube. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In Nicoletta Calzolari (Conference chair),

Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[175] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[176] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July 2020. Association for Computational Linguistics.

[177] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, 2019.

[178] Anastazia Zunic, Padraig Corcoran, and Irena Spasic. Sentiment analysis in health and well-being: Systematic review. *JMIR Med Inform*, 8(1):e16023, Jan 2020.

[179] Assia Soumeur, Mheni Mokdadi, Ahmed Guessoum, and Amina Daoud. Sentiment analysis of users on social networks: Overcoming the challenge of the loose usages of the algerian dialect. *Procedia computer science*, 142:26–37, 2018.

[180] Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273, 2017.

[181] Rosa María Montañes-Salas, Rafael del Hoyo-Alonso, and Rocío Aznar-Gimeno. From recurrency to attention in opinion analysis: Comparing rnn vs transformer models. In *IberLEF@SEPLN*, 2019.

[182] Lutfiye Seda Mut Altin, Àlex Bravo, and Horacio Saggion. Lastus/taln at tass 2019: Sentiment analysis for spanish language variants with neural networks. In *IberLEF@SEPLN*, 2019.

[183] Mohamed Amine Jerbi, Hadhemi Achour, and Emna Souissi. Sentiment analysis of code-switched tunisian dialect: Exploring rnn-based techniques. In *International Conference on Arabic Language Processing*, pages 122–131. Springer, 2019.

[184] Meysam Asgari-Chenaghlu, Narjes Nikzad-Khasmakhi, and Shervin Minaee. Covid-transformer: Detecting covid-19 trending topics on twitter using universal sentence encoder, 2020.

[185] Hui Yin, Shuiqiao Yang, and Jianxin Li. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. In Xiaochun Yang, Chang-Dong Wang, Md. Saiful Islam, and Zheng Zhang, editors, *Advanced Data Mining and Applications*, pages 610–623, Cham, 2020. Springer International Publishing.

[186] Amina Amara, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. Multilingual topic modeling for tracking covid-19 trends based on facebook data analysis. *Applied Intelligence*, 51(5):3052–3073, 2021.

[187] Jingyuan Yu, Yanqin Lu, and Juan Muñoz-Justicia. Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo. *International Journal of Environmental Research and Public Health*, 17(15):5414, 2020.

[188] Luis Carbonell Gironés. Geographical analysis of the opinion and influence of users on twitter during the coronavirus health crisis. Final project/degree, Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València, 2020.

[189] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[190] David Vilares and Carlos Gómez-Rodríguez. Grounding the semantics of part-of-day nouns worldwide using twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 123–128, 2018.

[191] Kevin P Scannell. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15, 2007.

[192] Justin Pinta. Lexical strata in loanword phonology: Spanish loans in guaraní. Master's thesis, The University of North Carolina at Chapel Hill, 2013.

[193] Lichan Hong, Gregorio Convertino, and Ed Chi. Language matters in twitter: A large scale study. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[194] Tim Kreutz and Walter Daelemans. Streaming language-specific Twitter data with optimal keywords. In *Proceedings of the 12th Web as Corpus Workshop*, pages 57–64, Marseille, France, May 2020. European Language Resources Association.

[195] Ron Artstein and Massimo Poesio. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 12 2008.

[196] Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online, June 2021. Association for Computational Linguistics.

[197] Luis Chiruzzo, Santiago Castro, and Aiala Rosá. HAHA 2019 dataset: A corpus for humor analysis in Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5106–5112, Marseille, France, May 2020. European Language Resources Association.

[198] World Health Organization (WHO). WHO statement regarding cluster of pneumonia cases in Wuhan, China. *World Health Organization (WHO)*, January 2020. Accessed: 2020-08-28.

[199] Purva Grover, Arpan Kumar Kar, Yogesh K Dwivedi, and Marijn Janssen. Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438–460, 2019.

[200] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973, 2011.

[201] Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace, editors. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[202] Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics.

[203] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[204] Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[205] Zubair Afzal, Vikrant Yadav, Olga Fedorova, Vaishnavi Kandala, Janneke van de Loo, Saber A. Akhondi, Pascal Coupet, and George Tsatsaronis. CORA: A deep active learning covid-19 relevancy algorithm to identify core scientific articles. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics.

[206] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[207] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalina, and Gerardo Chowell. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration, August 2020.

[208] Daniel Kerchner and Laura Wrubel. Coronavirus Tweet Ids, 2020.

[209] B. V. Barde and A. M. Bainwad. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750, 2017.

[210] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.

[211] Scott Deerwester, Susan Dumais, Thomas Landauer, George Furnas, and Laura Beck. Improving information-retrieval with latent semantic indexing. In *Proceedings of the ASIS annual meeting*, volume 25, pages 36–40. Information Today Inc 143 Old Marlton Pike, Medford, NJ 08055-8750, 1988.

[212] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 289–296, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[213] Suppawong Tuarob, Line C Pouchard, and C Lee Giles. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 239–248, 2013.

[214] Shibin Zhou, Kan Li, and Yushu Liu. Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, 2(4):398–409, 2009.

[215] Gu Yijun and Xia Tian. Study on keyword extraction with lda and textrank combination. *Data Analysis and Knowledge Discovery*, 30(7):41–47, 2014.

[216] Yang Gao, Yue Xu, and Yuefeng Li. Pattern-based topics for document modelling in information filtering. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1629–1642, 2014.

[217] Phu Pham, Phuc Do, and Chien DC Ta. W-pathsim: novel approach of weighted similarity measure in content-based heterogeneous information networks by applying lda topic modeling. In *Asian conference on intelligent information and database systems*, pages 539–549. Springer, 2018.

[218] David Andrzejewski and David Buttler. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 600–608, 2011.

[219] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[220] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA, 2006. Association for Computing Machinery.

[221] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225, May 2014.

[222] Ranganathan Chandrasekaran, Vikalp Mehta, Tejali Valkunde, and Evangelos Moustakas. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. *Journal of Medical Internet Research*, 22(10):e22624, 2020.

[223] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, 2012.

[224] Alaa Abd-Alrazaq, Dari Alhuwail, Mowafa Househ, Mounir Hamdi, and Zubair Shah. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016, 2020.

[225] Steven Loria. textblob documentation. *Release 0.16*, 2, 2020.

[226] Sakun Boon-Itt. A text-mining analysis of public perceptions and topic modeling during the covid-19 pandemic using twitter data. *JMIR public health and surveillance, JMIR Preprints. 30/06/2020:21978*, 2020.

[227] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[228] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *UMBC Faculty Collection*, 2020.

[229] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[230] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[231] Asbjørn Steinskog, Jonas Therkelsen, and Björn Gambäck. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st nordic conference on computational linguistics*, pages 77–86, 2017.

[232] Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer, 2010.

[233] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781, 2009.

[234] Stefan Wojcik and Adam Hughes. Sizing up twitter users. *PEW research center*, 24, 2019.

[235] Agencia EFE. La OMS pone en alerta a la red mundial de hospitales por un nuevo coronavirus en China. *www.efe.com*, January 2020.

[236] El Boletín. China pone en cuarentena a más de 30 millones de personas por el coronavirus. *elboletin.com*, January 2020.

[237] Beatriz Pérez. La OMS rectifica y declara la emergencia global por el coronavirus, January 2020.

[238] CatalunyaPress.es. Iberia suspende los vuelos a Shanghái por el coronavirus, 2020.

[239] Holly Ellyatt. Russia closes border with China to prevent spread of the coronavirus, January 2020.

[240] BBC News. Li Wenliang: Coronavirus kills Chinese whistleblower doctor. *BBC News*, February 2020.

[241] Pablo Linde. Sanidad confirma en La Gomera el primer caso de coronavirus en España. *El País*, February 2020.

[242] Marcos Pardeiro. El fracaso político del MWC: "No se va a suspender". "No cuelga de un hilo", 2020.

[243] Lucía Bohórquez and Oriol Güell. El segundo caso de coronavirus en España es un británico que se contagió en los Alpes. *El País*, February 2020.

[244] Paloma Almoguera. El coronavirus pone en jaque ahora a Japón y Corea del Sur. *El País*, February 2020.

[245] World Health Organization (WHO). Advice for the public on COVID-19 – World Health Organization, 2020.

[246] Hideyuki Sano. GLOBAL MARKETS-World stocks set for worst week since 2008 as virus fears grip markets. *Reuters*, February 2020.

[247] CNN. Medidas globales por el coronavirus: mantener distancia de un metro, cierre de escuelas y museos, evitar los besos y otras, March 2020.

[248] Carlos E. Cué. El Gobierno informa de que es la única autoridad en toda España, limita los desplazamientos y cierra comercios, March 2020.

[249] La Razón. Emotivo reconocimiento a los sanitarios en forma de aplausos desde los balcones, March 2020.

[250] José Polo. Coronavirus: La Zona Franca fabricará 100 respiradores diarios con impresoras 3D, March 2020.

[251] Cristian Fracassi. Charlotte valve, March 2020.

[252] RTVE.es. Los ERTE por la crisis del coronavirus suman más de 240.000, March 2020.

[253] Gestiona.es. Información para los afectados por ERTE debido al COVID19, March 2020.

[254] EFE/CMM. 400 guardias civiles de Castilla-La Mancha tienen Covid-19, según la AUGC, 2020.

[255] Alejandro Requeijo. La Policía y la Guardia Civil suman ya más de 400 positivos por coronavirus, March 2020.

[256] David Justo. España sigue la tendencia a la baja: 4.273 nuevos contagios por coronavirus y 637 muertes, April 2020.

[257] M.R.M. Un tigre del zoo de Nueva York tiene coronavirus, April 2020.

[258] La Vanguardia. Boris Johnson recibe el alta y continuará recuperándose de la Covid-19 en su casa, April 2020.

[259] Ana Soteras. COVID-19: 510 muertes en un día, la cifra más baja desde el 23 de marzo, 2020.

[260] Joan Safont Plumed. Muere el escritor chileno Luis Sepúlveda, a causa del coronavirus, 2020.

[261] elEconomista.es. Las medidas de distanciamiento social podrían extenderse hasta 2022 de manera intermitente - elEconomista.es, 2020.

[262] Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[263] Pieter Muysken. Code-switching and grammatical theory. *The bilingualism reader*, pages 280–297, 1995.

[264] Joseph Gafaranga and Maria-Carme Torras. Interactional otherness: Towards a redefinition of codeswitching. *International Journal of Bilingualism*, 6(1):1–22, 2002.

[265] Yaron Matras. *Language contact*. Cambridge University Press, 2020.

[266] Bruno Estigarribia. Guarani-spanish jopara mixing in a paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222, 2015.

[267] Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,"Visions for the Future of Machine Translation*, pages 1–9, 2006.

[268] Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France, May 2020. European Language Resources Association.

[269] Michael Gasser. Mainumby: un ayudante para la traducción castellano-guaraní. In *Libro del Tercer Seminario Internacional sobre Traducción, Terminología y Lenguas Minorizadas, Asunción, Paraguay, 19-21 julio, 2018*, volume 3, pages 328–355. Fundación Yvy Marae´Y, July 2018.

[270] Diego Manuel Maldonado, Rodrigo Villalba Barrientos, and Diego P Pinto-Roa. Eñe'ẽ: Sistema de reconocimiento automático del habla en guaraní. In *Simposio Argentino de Inteligencia Artificial (ASAI 2016)-JAIIO 45 (Tres de Febrero, 2016).*, 2016.

[271] Alexander James Rudnick. *Cross-Lingual Word Sense Disambiguation for Low-Resource Hybrid Machine Translation*. PhD thesis, Indiana University, 2018.

[272] Mika Hämäläinen. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345, 2019.

[273] Robert A Dooley. Léxico guarani, dialeto mbyá, com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística. esboço gramatical e referências. *Cuiabá: Summer Institute of Linguistics*, 2006.

[274] Guillaume Thomas. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France, August 2019. Association for Computational Linguistics.

[275] David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153, 2016.

[276] Dinkar Sitaram, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. Sentiment analysis of mixed language employing hindi-english code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 271–276. IEEE, 2015.

[277] Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for*

Under-Resourced Languages (CCURL), pages 202–210, Marseille, France, May 2020. European Language Resources association.

[278] Yanqing Chen and Steven Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[279] David Vilares, Carlos Gómez-Rodríguez, and Miguel A Alonso. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55, 2017.

[280] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[281] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[282] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[283] Jie Yang and Yue Zhang. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[284] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, 2014.

[285] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

[286] Yliana V Rodríguez. private communication - email, Oct 2021. `https://orcid.org/0000-0002-0311-2417` - The idea of this email-thread is to collaborate for analysis, study and discuss the border between the Guarani and the Jopara.

[287] Andrea Vilariño Ferreiro. The influence of social media in language change: Changes in vocabulary, 2018. (Bachelor's Thesis).

[288] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[289] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2020.

[290] Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Online, June 2021. Association for Computational Linguistics.

[291] Miguel Angel García Trillo, Aarón Estrella Gutiérrez, Alexander Gelbukh, Ana Paloma Peña Ortega, Antonio Reyes Pérez, Christian Efraín Maldonado Sifuentes, Damián Darío González Orozco, Diana Esperanza Vázquez González, Gemma Bel-Enguix, Gerardo Sierra Martínez, Grigori Sidorov, Hamlet Antonio García Zuñiga, Hermilo Santiago Benito, Horacio Rodríguez Hontoria, Iván Vladimir Meza Ruiz, Jason Efraín Ángel Gil, Jesús Manuel Mager Hois, Jorge García Flores, José Mateo Lino Cajero Velázquez, Leslie Denisse Soria Cruz, Margarita Abigail Mota Montoya, María Guadalupe Torres Felegrino, Noé Alejandro Castro Sánchez, Ricardo Ramos Aguilar, and Ricardo Ramos Ortega. *Procesamiento de lenguaje natural para las lenguas indígenas*. Number 1. Universidad Michoacana de San Nicolás de Hidalgo, 04 2021. `https://isbnmexico.indautor.cerlalc.org/catalogo.php?mode=detalle&nt=334970`.

[292] Yanina Borges, Florencia Mercant, and Luis Chiruzzo. Using guarani verbal morphology on guarani-spanish machine translation experiments. *Procesamiento del Lenguaje Natural*, 66(0):89–98, 2021.

[293] Anastasia Kuznetsova and Francis Tyers. A finite-state morphological analyser for Paraguayan Guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 81–89, Online, June 2021. Association for Computational Linguistics.

[294] Nicolás Giossa and Santiago Góngora. Construcción de recursos para traducción automática guaraní-español, 2021. (Bachelor's Thesis).

[295] Carlo Strapparava and Rada Mihalcea. Affect detection in texts. In *The Oxford Handbook of Affective Computing*, chapter 13. Oxford Library of Psychology, 2015.

[296] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992.

[297] Rada Mihalcea and Carlo Strapparava. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142, 2006.

[298] Zhongqing Wang, Shoushan Li, Fan Wu, Qingying Sun, and Guodong Zhou. Overview of nlpcc 2018 shared task 1: Emotion detection in code-switching text. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 429–433. Springer, 2018.

[299] Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. Humor detection in English-Hindi code-mixed social media content : Corpus and baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[300] Manikandan Ravikiran and Subbiah Annamalai. DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17, Kyiv, April 2021. Association for Computational Linguistics.

[301] Anshul Wadhawan and Akshita Aggarwal. Towards emotion recognition in Hindi-English code-mixed data: A transformer based approach. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 195–202, Online, April 2021. Association for Computational Linguistics.

[302] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online, November 2020. Association for Computational Linguistics.

[303] Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[304] Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online, April 2021. Association for Computational Linguistics.

[305] Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. In *Proceedings of the International Conference "Dialogue 2019"*, pages 333–339, Moscow, Russia, 2019. Computational Linguistics and Intellectual Technologies.

[306] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[307] David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. Bertinho: Galician bert representations. *Procesamiento del Lenguaje Natural*, 66(0):13–26, 2021.

[308] Giusepppe Attardi. Wikiextractor. `https://github.com/attardi/wikiextractor`, 2015.

[309] Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May 2020. European Language Resources Association.

[310] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[311] Daniel Myers and James W McGuffee. Choosing scrapy. *Journal of Computing Sciences in Colleges*, 31(1):83–89, 2015.

[312] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[313] Alexandre Devert. *matplotlib Plotting Cookbook*. Packt Publishing Ltd, 2014.

[314] Mitchell Hashimoto. *Vagrant: up and running: create and manage virtualized development environments*. O'Reilly Media, Inc., 2013.

[315] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014.

[316] Kyle Banker. *MongoDB in action*. Manning Publications Co., 2011.

[317] Dongjin Yu, Dengwei Xu, Dongjing Wang, and Zhiyong Ni. Hierarchical topic modeling of twitter data for online analytical processing. *IEEE Access*, 7:12373–12385, 2019.

[318] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2021.

[319] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[320] Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[321] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

[322] Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489, 2021.

[323] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[324] Yong Hu, Heyan Huang, Tian Lan, Xiaochi Wei, Yuxiang Nie, Jiarui Qi, Liner Yang, and Xian-Ling Mao. Multi-task learning for low-resource second language acquisition modeling. In Xin Wang, Rui Zhang, Young-Koo Lee, Le Sun, and Yang-Sae Moon, editors, *Web and Big Data*, pages 603–611, Cham, 2020. Springer International Publishing.

[325] Ahmed Magooda, Diane Litman, and Mohamed Elaraby. Exploring multitask learning for low-resource abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1652–1661, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[326] Antonio Usai, Marco Pironti, Monika Mital, and Chiraz Aouina Mejri. Knowledge discovery out of text data: a systematic review via text mining. *Journal of knowledge management*, 2018.

[327] Universidad Comunera. Resultados i-Data - Universidad Comunera. *Workshop Ciencia de Datos*, 2018. `https://www.ucom.edu.py/cienciadedatos/concurso-i-data/resultados/`.

[328] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.

[329] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

Appendices

Appendix A

Publications[1]

[1]We include the last manuscript submitted to the journal for its publication (and not directly the PDF-Portable Document Format file generated by the journal), following the 'Guide for the presentation and defense of doctoral theses of the University of Granada'. See the guide at `https://escuelaposgrado.ugr.es/doctorado/estudiantes/deposito/deposito_tesis?lang=en#__doku_you_must_submit_the_following_documentation`.

This appendix comprises publications that are part of this dissertation. These have been published in peer-reviewed conference and workshop proceedings, as well as in indexed journals.

A.1 A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor

- Agüero-Torales, M. M., Cobo, M. J., Herrera-Viedma, E., & López-Herrera, A. G. (2019). A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor. In Proceedings of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence (*Procedia Computer Science*, 162, 392-399). doi: `https://doi.org/10.1016/j.procs.2019.12.002`.
  - Status: Published.
  - Google Scholar metric[2] (h5 2019): index 84, median 111.
  - Subject Category: Engineering & Computer Science, Engineering & Computer Science (general). Ranking 13/20.
  - Scopus CiteScore[3] (2019): 2,5.
  - Subject Category: Computer Science, General Computer Science. Ranking 69/221.

A.1.1 Summary

Text mining uses statistical techniques on text for knowledge discovery, prediction, and classification [326]. We have performed a proof-of-concept contribution of a cloud-based tool used for sentiment analysis of English-language travelers' reviews on TripAdvisor.com. More particularly, about travelers' visiting restaurants in the Province of Granada (Spain). Our tool collects data from this website, performs an analysis of people and their comments, detects the polarity of comments (i.e. in positive, negative, and neutral classes), and also allows all operations such as data collection, preprocessing, analysis and visualizations can be performed in the same application. Note that we chose the TripAdvisor.com platform because it provides a large amount of raw data.

This paper conducted a sentiment analysis study covering the analysis of customer reviews in the tourism and restaurant industry. Specifically, the tool presented for sentiment analysis was applied to an analysis of TripAdvisor reviews focused on the restaurant industry. Our tool performs some textual data processing steps, such as small text-parsing, tokenization of reviews in sentences, etc. It performs sentiment analysis with VADER [221] on the raw reviews written in English. Lastly, the tool's final evaluation combines the sentiment analysis based on the text of the reviews and the direct comments from individuals.

In a first version of this article, we presented the abstract as a poster called *Gastrominer. Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor: Caso de Estudio sobre Restaurantes de la Provincia de Granada* (in Spanish) [92], in an applied data science contest organized by the Universidad Comunera,[4] 'i-Data', in Asunción (Paraguay) in October 2018, located in the 'Jornadas Científicas de Ciencia de Datos', achieving first place [327]. See the poster at section A.2.

[2]See Google Scholar metrics at `https://scholar.google.com/intl/en/scholar/metrics.html#metrics`.

[3]View Scopus metric methodology at `https://service.elsevier.com/app/answers/detail/a_id/14880/supporthub/scopus/`.

[4]`https://www.ucom.digital/`

Procedia Computer Science 00 (2019) 000–000

Information Technology and Quantitative Management (ITQM 2019)

A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor

M.M. Agüero-Torales[a,1], M.J. Cobo[b], E. Herrera-Viedma[a], A.G. López-Herrera[a,1]

[a]*Dept. of Computer Science and Artificial Intelligence, University of Granada, Calle Daniel Saucedo Aranda, s/n, 18071, Granada, Spain*
[b]*Dept. Computer Science and Engineering, University of Cádiz, Avenida Ramón Puyol, 11202, Algeciras, Cádiz, Spain.*

Abstract

The tourism industry has been promoting its products and services based on the reviews that people often write on travel websites like TripAdvisor.com, Booking.com and other platforms like these. These reviews have a profound effect on the decision making process when evaluating which places to visit, such as which restaurants to book, etc.

In this contribution is presented a cloud based software tool for the massive analysis of this social media data (TripAdvisor.com). The main characteristics of the tool developed are: i) the ability to aggregate data obtained from social media; ii) the possibility of carrying out combined analyses of both people and comments; iii) the ability to detect the sense (positive, negative or neutral) in which the comments rotate, quantifying the degree to which they are positive or negative, as well as predicting behaviour patterns from this information; and iv) the ease of doing everything in the same application (data downloading, pre-processing, analysis and visualisation).

As a test and validation case, more than 33.500 revisions written in English on restaurants in the Province of Granada (Spain) were analysed.

Selection and/or peer-review under responsibility of ITQM2019.

Keywords: opinion mining, sentiment analysis, TripAdvisor, software tool, cloud

1. Introduction

With the advent of new ICT (Information and Communication Technologies), supported by Web 2.0, millions of people create billions of connections through the media. Each click and each keystroke creates relationships that together form a vast social network.

Users of social communication tools (email, blogs, microblogs, wikis, etc.) send fervent personal or public messages, vigorously publish opinions about a product, a person or an event, or contribute altruistically and disinterestedly to the community of knowledge to make collaborations, promote cultural heritage, or advance the cultural in the development of some product or idea. Passionate about social networks create and share (texts, images, videos, links, etc.) and value or recommend products, people and services providing help to others (whether they are neighbors or live in each other's home). extreme of the world), and expressing their creativity (for example, photos on Flickr.com or Instagram.com; videos on YouTube.com or Vimeo.com; etc.);

[1]Corresponding author. Tel.: +34-958-248557; fax: +34-958-243317.
E-mail address: lopez-herrera@decsai.ugr.es.

thus contributing to Intelligence Web Collective. The result of all this is vast and tremendously complex networks of connections that relate people to each other, and to documents, locations, concepts, and all class of objects (mostly digital).

New opinion mining tools are now more than ever needed to collect, analyse, visualise, and generate in-depth knowledge (in the form of insights) from connection sets made up of millions of messages, links, entries, edits, photo and video updates, reviews, and product recommendations. These tools could help organisations in several ways: to know what is said (whether good or not so good) about products, services, departments of the organisation, or even the entity itself, in what sense the opinions of customers or potential consumers. To know how organisations could improve their image, products and services.

Sentiment analysis is the task of identifying and classifying the sentiments and opinions expressed in a text to understand the attitude towards a product, theme, service, etc. in particular [1]. Thus, the objective of this contribution is the presentation of a tool for opinion mining and sentiment analysis. The tool is cloud-based and focuses on the gastronomy sector and will feed from opinions published on the platform TripAdvisor.com. In order to test and validate the tool, an analysis will be made of the gastronomic context of the Province of Granada.

The rest of the contribution is structured as follows. Section Methodology summarises the main steps carried out by the tool. The proposal section displays the cloud-based technology used by the tool. Section Validation shows some of the results that the tool is able to produce. Finally, some conclusions are drawn.

2. Methodology

The methodology was divided in four steps or stages. The first was the data collection stage, next to the text preparation stage, the sentiment analysis stage with a simple rule-based model, and finally, the representation stage (evaluation and visualisation of the results). See Figure 1 (left).

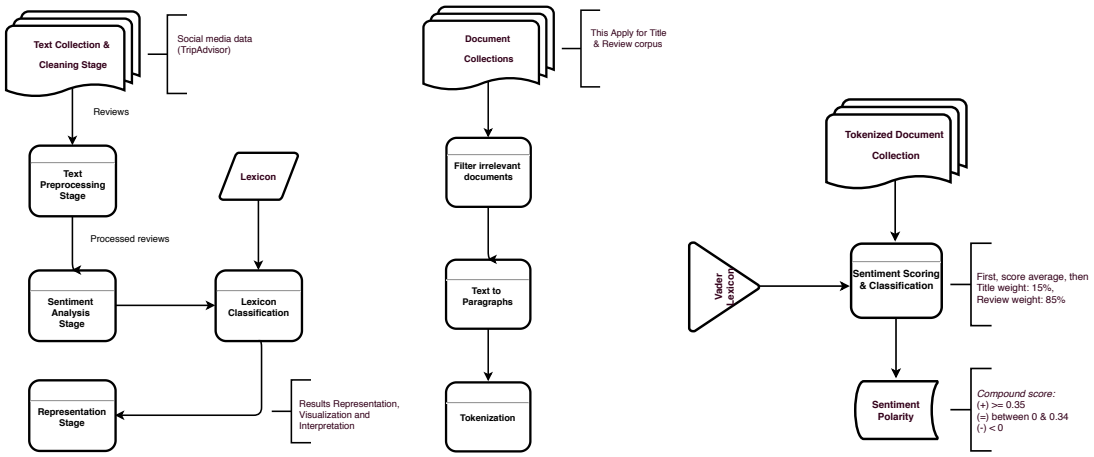


Fig. 1. Flow of the methodology. The full methodology stages (left). The text preparation stage (center). The sentiment analysis stage (right).

2.1. Data collection

We scrape TripAdvisor web-pages of all Restaurants in the Granada Province of Spain: included Bars & Pubs, Speciality Food Market, Quick Bites, Dessert, Coffee & Tea and Bakeries. According with Liu proposal [2], an opinion is a quadruple, composed by: sentiment orientation (s), sentiment target (g), opinion holder (h) and time (t). We contemplate taking some restaurants fields based on this proposal and saving them as a collection of documents (considering the following model that maps restaurants and multiple opinion relationships) [3]:

- name
- URL

- rating
- reviews collections
  - title
  - review
  - user-name
  - date
  - rating
- address
  - street
  - postal code
  - locality

As in a typical data collection stage [4], the HTML text is parsed, then scraped on TripAdvisor.com: the restaurant search URL of the Province of Granada is received, processed and navigated, retrieving all restaurant reviews (and some important data of the establishment) and, finally, reviews are recorded in a database. All reviews collected were written in English and we consider user ratings with one and two bubbles, as negative, three, as neutral, and four and five, as positive. We collected 33,594 user comments written until September 2017.

2.2. Text preparation

For text preparation stage the restaurants without reviews was discarded. The review title and the review was tokenize in paragraphs and sentences, and the text was conserved as are it, stop-words was removed but not special characters as emoticons or any others. Figure 1 (center) illustrates the text preparation stage.

In most cases on the TripAdvisor reviews, it is rare to see links next to the text, maybe emoticons, but for the sentiment analysis we use a emoticons corpus to treat it. Basically, a text analysis is performed without a very careful data cleansing strategy, based on valence[1], where the intensity of the feeling is taken into account. In turn, this type of sentiment analysis is based on word lexicons (and emoticons). By using this approach, each word in the lexicon is classified as to whether it is positive or negative, almost always how positive or negative it is [6].

2.3. Sentiment analysis

The different levels of sentiment analysis are the document level, to classify whether a whole opinion document expresses a positive or negative sentiment [7] [8], the sentence level, to determine whether each sentence expresses a positive, negative, or neutral opinion. Finally, the aspect level, to discover the target of opinion: what people like and dislike exactly [2].

In Figure 1 (right) you can see the stage of sentiment analysis, its check if any of the words in the sentences are present in the lexicon, the input text produces an output of four feeling metrics from these word ratings, which contains the proportion of negative, positive and neutral words of the given text, and a compound value, with values between -1 and 1. The compound value determines the polarity of the text, is a sum of all the qualifications of the lexicon applied [6].

2.4. Representation and web interface

The representation stage exposes what is interesting about the data, which covers the results representation and reporting, their visualisation and interpretation to understand them [1] [9].

This stage implies a responsive design for the web interface with a menu of options, also a list of restaurants. Figure 2 (top) illustrates the home screen with statistics.

In Figure 2 (bottom-left), the restaurant details screen, the user can see the polarity of sentiments of all the reviews and one-by-one, such as an alternative weighting factor to TripAdvisor bubbles, also a link to TripAdvisor.com to corroborate the sources, the same capabilities have the user details screen (see Figure 2 (bottom-right)), which shows the overall polarity of the user, but restaurant screen has a link to Google Maps to see more extra comments and information.

[1] Emotional valence is a psychology term refers to the intrinsic attractiveness (good-ness, positive valence) or averseness (bad-ness, negative valence) of an event, object, or situation, especially used when discussing emotions [5].

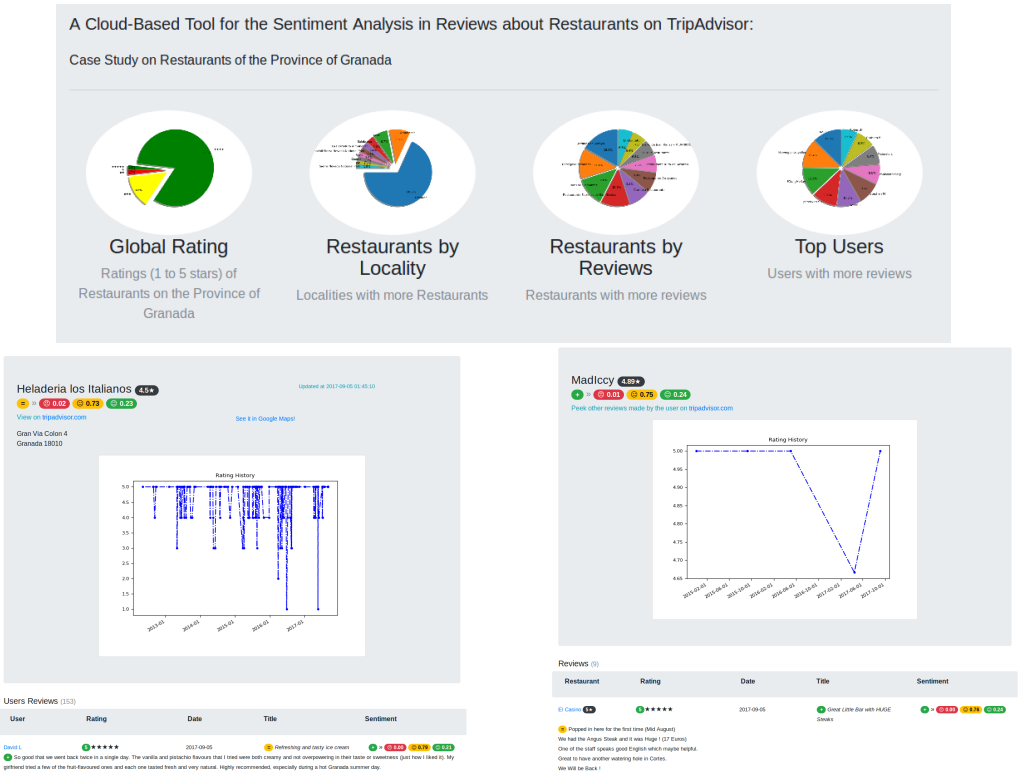


Fig. 2. Kind of analysis possibilities: general, restaurants, users.

Other important features are the on-demand TripAdvisor scraper based on an HTTP server that can control scraping through HTTP requests, the restaurants simple search engine, the user manager for authentication and registration, a REST API from the restaurant collection to load more data or download existent ones, and the calculation of the sentiment polarity on request for a particular analysis.

2.5. *Software tool*

Modern architectures have as a common general objective to seek consistency in the speed of response to the user, the use of own resources combined with third-party resources, and are based on agile development methodologies and continuous integration and continuous deployment; unlike a monolithic one, which is not suitable for modern applications because it has a single client running the user interface and a single server (replicated or not) running all the components of that application, neither because of its scalability characteristics, nor for the distribution of tasks and data between different parts of a development team [10].

To create the software tool for mass analysis of social media data, container technology was used. Containers are virtualized at the operating system level, while hypervisor-based solutions are virtualized at the hardware level, therefore, for a container, resource utilization is much more efficient, isolated and cheaper (even both technologies can complement each other for this reason) [11].

In addition of the container technology, a virtualized web environment were be created accessible from any device from a browser, as it does computing in the cloud for modern applications (see Figure 3). Then, instead of having a single service, by applying a micro-service view, each container becomes a service by itself, and the services communicate with each other through calls, gaining remarkable scalability.

Fig. 3. Web architecture and deployment diagram [12].

3. Proposal

The software tool were developed with a Python stack [1], support by Scrapy, a web crawling framework for web scraping [13], NLTK (Natural Language ToolKit), a toolkit for Natural Language Processing (NLP) provides interfaces to over 50 corpora and lexical resources , along with a suite of text processing libraries (e.g., classification, tokenization) [14], Matplotlib, a 2D plotting library for data visualisation [15], and Django[2], a potent web framework with many plugins that encourages rapid development.

The web environment relies in the virtualisation technology, focus on automation with Vagrant, a tool for building and managing virtual machine environments in a easy-to-use and single workflow [16], VirtualBox[3], a cross-platform virtualization application, a robust Docker stack, for harnessing the benefits of containerization for a focused purpose (i.e., the lightweight packaging and deployment of applications) [11], and for the data persistence, MongoDB, a general purpose, document-based, distributed database [3], where each restaurant is a document. See the Figure 4.

The sentiment analysis stage relied in VADER (Valence Aware Dictionary for sEntiment Reasoning), a lexicon and rule-based sentiment analysis tool with an overall precision of 99% for tweets sentiment classification, this tool is embedded in NLTK as a module, which proportionate other tools for text preparation and cleaning, all-in-one, but VADER no need many text preparation, it incorporate word-order sensitive relationships between terms for punctuation (e.g.,!), capitalization (e.g., ALL-CAPS), degree modifiers, support for the contrastive conjunction ”but”, tri-grams preceding a sentiment-laden lexical feature, or emoticons, etc. [6]. In this pipeline stage, the title and the text of the reviews is prepare and send to VADER as a sentence, the title has a weight of 15% and the review has a 85% for determine the overall polarity of the review. In a paragraph, each sentence is calculated with VADER and are calculated one-to-one for take the average of each one. This average of the compound scores (between -1 and 1) determine the final polarity, that is positive, if average equals or is major to 0.35, neutral, if average is between 0 and 0.34, and negative, if minor to 0.

4. Validation

In order to test the tool, some analyses were carried out. For example, ten best restaurants in the Province of Granada were brought together according to the ratings (date: September 13th, 2017) of TripAdvisor users and the number of revisions (see Table 1). According to this ranking, in that date the best restaurant was ”El Mercader” (Granada) and the tenth was ”Restaurante Las Chimeneas” (Mairena, Nevada). For these restaurants the sentiment polarity was individually calculated by the software tool, and the results can be seen in Figure 5. It can be seen how the restaurant ”Duran Barista” was the one that added more positivity among the reviews that were made about it. There were hardly any negative ratings. Almost the majority of the assessments were neutral.

By way of illustration, some other results that the tool is capable of producing are shown in the Figure 6. The overall rating on restaurants in the Province of Granada is shown at the top-left. Restaurants with the highest

[2]https://www.djangoproject.com/foundation/
[3]https://www.virtualbox.org/wiki/VirtualBox

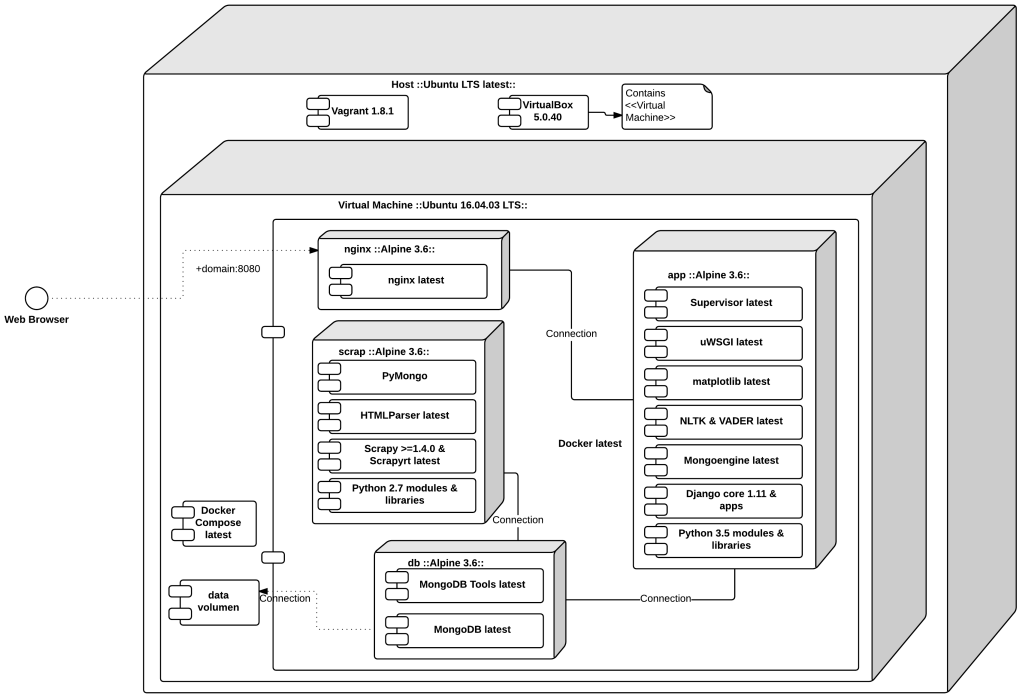


Fig. 4. General architecture deployed.

Table 1. The ten best restaurants in the Province of Granada by TripAdvisor user rating and number of reviews. Date: September 13th, 2017.

	Restaurant	Stars	Compound	Negative	Neutral	Positive	Locality	Reviews
1	El Mercader	5	0,45	0,01	0,67	0,31	Granada	152
2	Restaurante Casa Piolas	5	0,39	0,01	0,73	0,26	Algarinejo	70
3	Duran Barista	5	0,46	0,02	0,65	0,34	Granada	40
4	Dulcimena Coffee & Go	5	0,44	0,02	0,68	0,30	Granada	32
5	D'eti Coffee And Cake	5	0,51	0,01	0,66	0,32	Granada	32
6	Vega - Foodie Bar	5	0,48	0,01	0,67	0,30	Granada	28
7	La Huella Gastrobar	5	0,48	0,01	0,64	0,33	Algarinejo	27
8	Colagallo Craft Beers & Cocktails	5	0,51	0,01	0,65	0,32	Granada	24
9	Eco De.leite	5	0,42	0,04	0,68	0,28	Granada	14
10	Restaurante Las Chimeneas	5	0,46	0,02	0,68	0,28	Mairena	13

number of reviews in the Province of Granada are shown at the top-right. The locations with more establishments in the Province of Granada can be seen at the bottom-left. And finally, the bottom-right of the Figure 6 shows the Users with the highest number of restaurant reviews.

5. Conclusion

TripAdvisor is a social media par excellence to share opinions on sites related to travel, such as gastronomic establishments, that's why it has thousands of reviews a day.

With this in mind, the purpose of this contribution is to present a tool for analyzing the sentiments of gastronomic establishments. The tool has been developed using cloud-based technologies and allows all the necessary steps to carry out this type of analysis, from downloading data to displaying results, without forgetting all the intermediate stages of pre-processing, cleaning, data preparation, dimensionality reduction, etc.

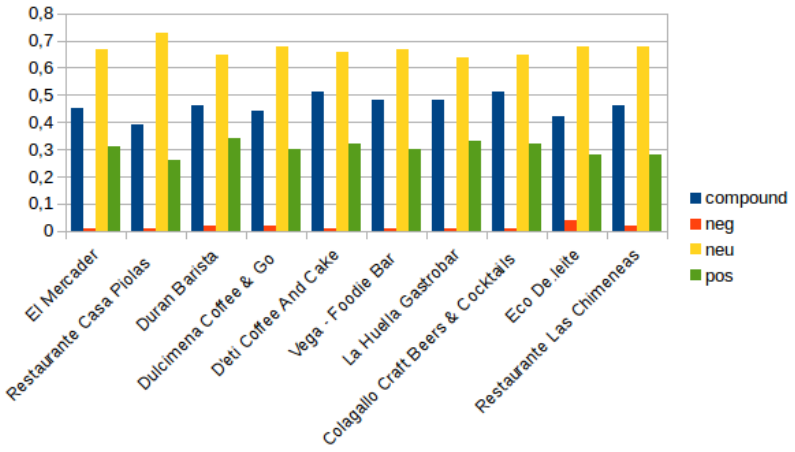


Fig. 5. Results about the ten first restaurants of Granada.

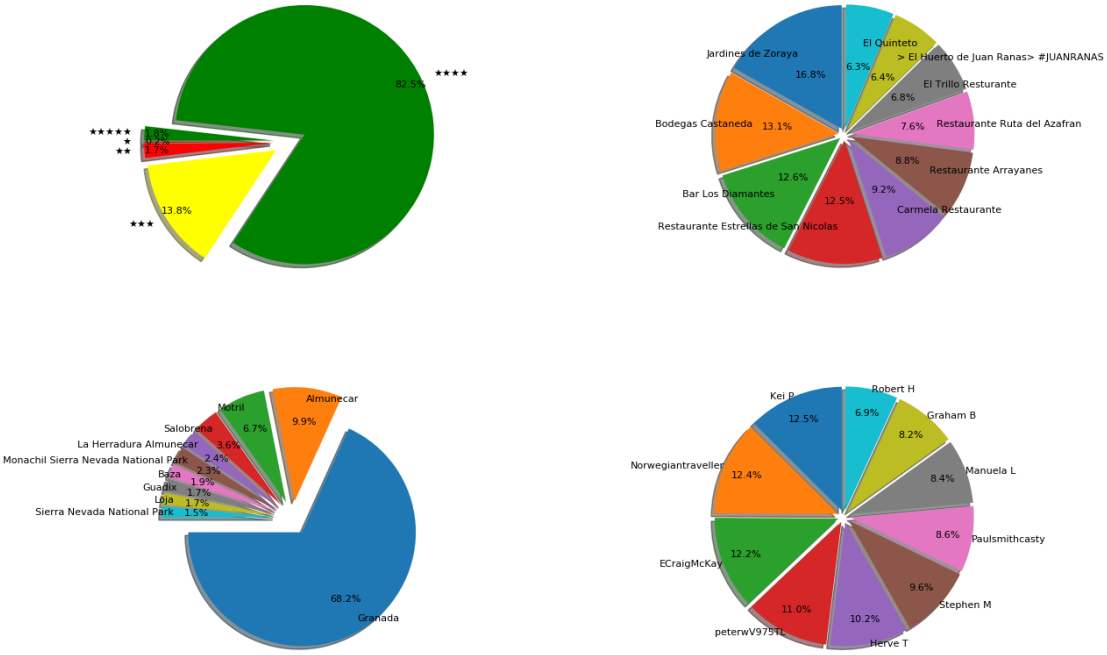


Fig. 6. Some other results. Overall rating on restaurants in the Province of Granada (top-left). Restaurants with the highest number of reviews in the Province of Granada (top-right). Locations with more establishments in the Province of Granada (bottom-left). Users with the highest number of restaurant reviews in the Province of Granada (bottom-right).

The tool has been successfully validated in the gastronomic context of the Province of Granada (Spain). Some results have been shown in this contribution.

As future works we plan the improvement of some components of the tool, such as the integration with other data sources of specific purpose as Booking.com, and others of general purpose as Twitter, Instagram, YouTube, etc.

We also plan to delve deeper into online sentiment analysis methods such as MeaningCloud[4], Indico.Io[5] or Google Cloud[6], as well as explore other approaches such as aspect analysis.

References

[1] M. Bonzanini, Mastering social media mining with Python, Packt Publishing Ltd, 2016.
[2] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.
[3] K. Banker, MongoDB in action, Manning Publications Co., 2011.
[4] R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, " O'Reilly Media, Inc.", 2018.
[5] N. H. Frijda, The Emotions, Cambridge University Press, 1986.
[6] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eight International AAAI conference on weblogs and social media, 2014.
[7] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79–86.
[8] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
[9] D. A. Keim, Information visualization and visual data mining, IEEE transactions on Visualization and Computer Graphics 8 (1) (2002) 1–8.
[10] J. J. Merelo, University of Granada, Lecture notes in Cloud Computing: Arquitecturas software para la nube (September 2019). URL `https://jj.github.io/CC/`
[11] D. Merkel, Docker: lightweight linux containers for consistent development and deployment, Linux Journal 2014 (239) (2014) 2.
[12] J. M. Guirao, University of Granada, Lecture notes in Sistemas Software Basados en Web: Preliminares (September 2017). URL `https://swad.ugr.es/es`
[13] D. Myers, J. W. McGuffee, Choosing scrapy, Journal of Computing Sciences in Colleges 31 (1) (2015) 83–89.
[14] E. Loper, S. Bird, Nltk: the natural language toolkit, arXiv preprint cs/0205028.
[15] A. Devert, matplotlib Plotting Cookbook, Packt Publishing Ltd, 2014.
[16] M. Hashimoto, Vagrant: up and running: create and manage virtualized development environments, " O'Reilly Media, Inc.", 2013.

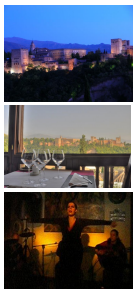
[4] https://www.meaningcloud.com/developer/sentiment-analysis
[5] https://www.indico.io/blog/docs/indico-api
[6] https://cloud.google.com/natural-language/docs/

A.2 GASTRO-MINER - Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor: Caso de Estudio sobre Restaurantes de la Provincia de Granada

- Agüero-Torales, M. M., López-Herrera, A. G. & Cobo, M. J. (2018). GASTRO-MINER - Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor: Caso de Estudio sobre Restaurantes de la Provincia de Granada. In LIBRO DE RESÚMENES - I Jornadas Científicas en CIENCIA DE DATOS, (pp. 34). Universidad Comunera. First place in the 'i-Data' contest. doi: http://dx.doi.org/10.13140/RG.2.2.33136.71688.[5]

  – Status: Published.

[5] https://www.ucom.edu.py/wp-content/uploads/2019/08/Jornadas-Cientificas-en-Ciencia-de-Datos-UCOM.pdf#page=34



UNIVERSIDAD DE GRANADA
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACION E INTELIGENCIA ARTIFICIAL

GASTRO-MINER

Una Herramienta Basada en la Nube para el Análisis de Sentimientos en Opiniones sobre Restaurantes en TripAdvisor: Caso de Estudio sobre Restaurantes de la Provincia de Granada

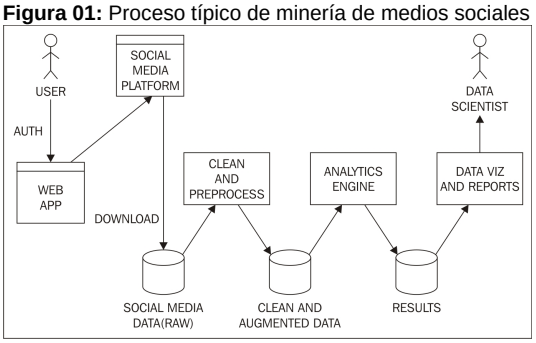
Aguero-Torales, M. M. [1, a]; Lopez-Herrera, A.G. [1, b]; Cobo, M.J. [2]
[1] Universidad de Granada, Granada, España, [a] maguero@correo.ugr.es; [b] lopez-herrera@decsai.ugr.es; [2] Universidad de Cádiz, Cádiz, España, manueljesus.cobo@uca.es

1. INTRODUCCIÓN

La industria del turismo ha estado promoviendo sus productos y servicios basados en las revisiones que las personas a menudo escriben en los sitios web de viajes como tripadvisor. Estas revisiones tienen un efecto profundo en el proceso de toma de decisiones cuando se evalúan qué lugares visitar, como en cuáles restaurantes reservar.

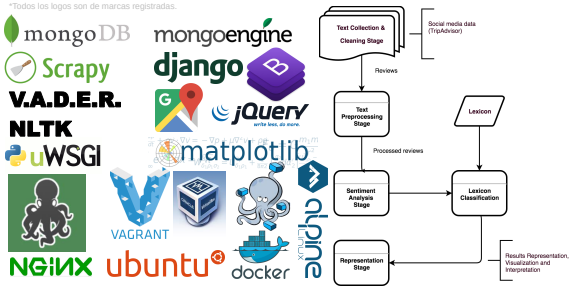
Desarrollar una herramienta software para el análisis masivo de datos procedentes de medios sociales.

Figura 01: Proceso típico de minería de medios sociales



Fuente: De *Mastering Social Media Mining with Python* por Bonzanini, M., 2016.

2. MATERIAL Y MÉTODOS



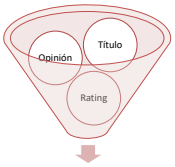
3. RESULTADOS Y DISCUSIÓN

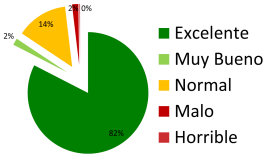
Se desarrolla una herramienta software para el análisis masivo de datos procedentes de medios sociales tripadvisor de la Provincia de Granada en (se reunieron más de 33.500 revisiones escritas en Inglés sobre restaurantes), cuyas características son:

- la capacidad de agregar datos obtenidos de medios sociales;
- la posibilidad de realizar análisis combinados tanto de personas, como de comentarios;
- la facultad de detectar el sentido (☺, ☹ o 😐) en el que giran los comentarios, cuantificando el grado en el que son ✚ o ▬, así como predecir patrones de comportamiento a partir de dicha información;
- la facilidad de realizar todo en una misma aplicación (descarga de datos, preprocesamiento, análisis y visualización).

4. CONCLUSIÓN

"Jardines de Zoraya", restaurante con más opiniones.





Descargar ahora la presentación **bit.ly/gastro-miner**

*Todas las fotografías son de usuarios de tripadvisor.com.

A.3 Discovering topics in Twitter about the COVID-19 outbreak in Spain

- Agüero-Torales, M. M., Vilares, D., & López-Herrera, A. G. (2021). Discovering topics in Twitter about the COVID-19 outbreak in Spain. *Procesamiento del Lenguaje Natural*, 66, 177-190. URL:`http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6333`.
  - Status: Published.
  - Citation Indicator (JCI 2020): 0,18.
  - Subject Category: Linguistics. Ranking 236/262 (Q4).
  - Scopus CiteScore (2020): 1,0.
  - Subject Category: Arts and Humanities, Language and Linguistics. Ranking 241/879.
  - Subject Category: Social Sciences, Linguistics and Language. Ranking 265/935.
  - Subject Category: Computer Science, Computer Science Applications. Ranking 553/693.

A.3.1 Summary

In this work, we automatically extracted topics of Spanish (Castilian) tweets from different periods of the first wave of the COVID-19 quarantine. This article presented a topic modeling approach to capture what Twitter users from Spain were discussing during the beginning of the COVID-19 pandemic. We presented an analysis of these topics, in a current study, due to the use of data from the pandemic we are currently suffering since 2020.

We used NLP strategies to perform an exhaustive qualitative analysis of these tweets and the evolution of the topics. In particular, we considered three different periods: pre-crisis, pandemic outbreak, and lockdown. The work was divided into the following phases: (i) collection of COVID-19 related tweets using keywords; (ii) language identification of the tweets (via several tools) and geolocation (via lists of Spanish cities and regions); (iii) tweet preprocessing; and (iv) topic modeling using Latent Dirichlet Allocation (LDA) [230], which is a standard and has proved to be robust for many tasks (see section 4.2). Finally, we presented a qualitative results analysis based on matching the extracted topics with newspaper news. To represent the topics, we used the traditional generative route, and also we introduced a discriminative route, extracting the most salient keywords and sentences. Thus, we provided a methodology and a domain-specific geolocalized public dataset.

We also performed a small quantitative evaluation framework based on a human evaluation, since it was difficult to find many works that perform topic modeling in Spanish about COVID-19 (and especially the one spoken in Spain).

Discovering topics in Twitter about the COVID-19 outbreak in Spain

Descubriendo temas en Twitter sobre el brote del COVID-19 en España

Marvin M. Agüero-Torales[1], David Vilares[2], Antonio G. López-Herrera[1]
[1]University of Granada, Spain
[2]Universidade da Coruña, CITIC, Spain
maguero@correo.ugr.es, david.vilares@udc.es, lopez-herrera@decsai.ugr.es

Abstract: In this work, we apply topic modeling to study what users have been discussing in Twitter during the beginning of the COVID-19 pandemic. More particularly, we explore the period of time that includes three differentiated phases of the COVID-19 crisis in Spain: the pre-crisis time, the outbreak, and the beginning of the lockdown. To do so, we first collect a large corpus of Spanish tweets and clean them. Then, we cluster the tweets into topics using a Latent Dirichlet Allocation model, and define generative and discriminative routes to later extract the most relevant keywords and sentences for each topic. Finally, we provide an exhaustive qualitative analysis about how such topics correspond to the situation in Spain at different stages of the crisis.
Keywords: COVID-19, Twitter, social networks, topic modeling.

Resumen: En este trabajo, analizamos lo que los usuarios han estado discutiendo en Twitter durante el comienzo de la pandemia causada por el COVID-19. Concretamente, analizamos tres fases diferenciadas de la crisis del COVID-19 en España: el propio tiempo de pre-crisis, el estallido de la enfermedad y el confinamiento. Para llevar esto a cabo, primero recolectamos una gran cantidad de tuits que son preprocesados. A continuación, agrupamos los tuits en distintas temáticas usando un modelo de Latent Dirichlet Allocation, y definimos estrategias generativas y discriminativas para extraer las palabras clave y oraciones más representativas para cada tema. Finalmente, incluimos un exhaustivo análisis cualitativo sobre dichos temas, y cómo estos se corresponden con distintas problemáticas surgidas en España en distintos momentos de la crisis.
Palabras clave: COVID-19, Twitter, redes sociales, modelado de temas.

1 Introduction

The outbreak of the SARS-CoV-2 virus and the global spread of the COVID-19 disease has encouraged people and organizations to express their opinion, discuss topics and warn about the evolution of the pandemic in social media platforms such as Twitter.

Unlike previous occasions, such as SARS-CoV in 2002 (World Health Organization (WHO), 2020b), where social media still were in an embryonic state and natural language processing (NLP) still had limited practical applications; we are now in a situation where users generate a vast amount of written content, that can be analyzed by automatic tools to discover the topics societies care about, and their sentiment. This has been already the case for some precedent events or catastrophes in recent years, such as the 2016 US political elections (Grover et al., 2019) or some natural disasters, such as the 2011 East Japan Earthquake (Neubig et al., 2011).

In relation to the COVID-19 pandemic, a few specific NLP workshops (Verspoor et al., 2020b; Verspoor et al., 2020a) have already attempted to highlight how NLP can be used to respond to situations like the current one; addressing a number of challenges that include mining scientific literature and social media analysis, among many others (Wang et al., 2020; Kleinberg, van der Vegt, and Mozes, 2020; Afzal et al., 2020). With research purposes, there has been also efforts on releasing NLP datasets discussing COVID-19 topics (Chen, Lerman, and Fer-

rara, 2020; Banda et al., 2020; Kerchner and Wrubel, 2020). In this context, the area of topic modeling has not been a stranger to this problem, and a number of authors have showed the options that clustering online posts such as tweets or Facebook messages can offer to monitor and evaluate the evolution of the pandemic through time (Asgari-Chenaghlu, Nikzad-Khasmakhi, and Minaee, 2020; Yin, Yang, and Li, 2020; Amara, Taieb, and Aouicha, 2020).

Contribution In this work, we also focus on the possibilities of performing effective and representative topic modeling over a large set of Spanish tweets. More particularly, we first collect a few millions tweets about COVID-19, mostly between 1 January to 20 April of 2020. Then, we apply latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003) to compute relevant topics in an unsupervised way, and obtain meaningful keywords and sentences through generative and discriminative routes. Finally, we provide an analysis to shed some light about the quality of the extracted topics, and how faithfully they represent what was happening in the Spanish society at different moments of the pandemic.

2 Related work

In what follows, we review topic modeling and NLP research related to COVID-19.

2.1 Topic modeling

In topic modeling, a topic is often viewed as a pattern of co-occurring words that can be exploited to cluster together documents from a large collection (Barde and Bainwad, 2017). Among methods for topic modeling we can find approaches such as the Vector Space Model (VSM) (Salton, Wong, and Yang, 1975), Latent Semantic Indexing (LSI) (Deerwester, 1988), Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) or lda2vec (Moody, 2016). Related to this, one of the most well-known, standardized and widely-used methods is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). More particularly, LDA is an unsupervised clustering approach where documents can belong to multiple topics, and where each topic is a mix of words, which can be shared among topics too.

The applications of these topic modeling approaches are many and include areas such as tag recommendation (Tuarob, Pouchard, and Giles, 2013), text categorization (Zhou, Li, and Liu, 2009), keyword extraction (Yijun and Tian, 2014), information filtering (Gao, Xu, and Li, 2014), similarity search in the fields of text mining (Pham, Do, and Ta, 2018), and information retrieval (Andrzejewski and Buttler, 2011).

2.2 Text Mining on English COVID-19 related tweets

With the COVID-19 outbreak, different authors have tried to apply topic modeling and text mining techniques to help analyze and monitor the situation of the pandemic, with a great focus on English messages. For instance, Asgari-Chenaghlu, Nikzad-Khasmakhi, and Minaee (2020) analyzed English tweets and detected the trending topics and major concerns of people with respect to COVID-19, by proposing a model based on the Universal Sentence Encoder (Cer et al., 2018). The model first derives a semantic representation and similarity of tweets and, over those similar tweets, it applies text summarization techniques to provide a summary of different clusters. In a related line, Yin, Yang, and Li (2020) proposed a framework to analyze the topic and sentiment changes in society over time due to the COVID-19, using Twitter to collect the source data. More specifically, they used a dynamic LDA for topic modeling over fixed time intervals (Blei and Lafferty, 2006), and VADER for sentiment analysis (Hutto and Gilbert, 2014). Chandrasekaran et al. (2020) examined the key topics among 13.9M English tweets about COVID-19, dealing with areas such as economy and markets, spread and growth in cases, treatment and recovery, impact on the healthcare sector, and governments response. They explored the trends and variations, and how those key topics, and associated sentiments changed over a period of time of 17 weeks, between 1 January 2020 and 9 May 2020. More particularly, they used guided LDA for topic modeling (Jagarlamudi, Daumé III, and Udupa, 2012), an LDA-variant where the model is guided to learn topics that are of specific interest, using priors in the form of seed words, and again VADER for sentiment analysis.

Also, Abd-Alrazaq et al. (2020) use LDA to detect topics such as the origin of the virus and its impact on people and coun-

tries, analyzing 2.8M English tweets. In addition, they performed sentiment analysis with `TextBlob` (Loria, 2020) and extracted some social network statistics for each topic, such as the number of followers, the number of likes of tweets, the number of retweets, the user mentions, or the link sharing, calculating the interaction rate per topic. At a smaller scale (100K English tweets) and considering only the pre-crisis lockdown period (from 12 December 2019 to 9 March 2020); Boon-Itt (2020) presented a work to understand public perceptions of the trends of the COVID-19 pre-pandemic time. The analysis included time series, sentiment analysis and emotional tendency using the NRC sentiment lexicon (Mohammad and Turney, 2013), as well as topic modeling using LDA.

2.3 Text Mining on Spanish and Multilingual COVID-19 related tweets

As usual in NLP, most of early efforts to monitor COVID-19 user-generated texts have focused on English. However, some work is already available for the Spanish language. For instance, Yu, Lu, and Muñoz-Justicia (2020) compare the news updates of two of the main Spanish newspapers Twitter accounts, *El País* and *El Mundo*, during the pandemic; applying topic modeling and network analysis methods. They identified eight news frames for each newspaper and split it in three clusters: the pre-crisis period (from 19 February to 14 March of 2020), the lockdown period (from 14 March to 11 May of 2020) and the recovery period (from 11 May to 3 June of 2020). Their goal was to understand how the Spanish news media covered the public health crisis in Twitter.

Besides, Carbonell Gironés (2020) proposed a geographical analysis of the opinion and influence of users in Twitter during the covid health crisis, considering tweets written in English and Spanish, and using LDA topic modeling. The first part of the study was a general approach to the analysis of the topics of US and UK users. The second part was an analysis of the interests of Twitter users in Spain during the confinement period (from 14 March to 22 July of 2020). To geolocate the tweets, they performed a country-level search for the English dataset, and a city or province-level search for the Spanish dataset; looking in both cases for any geographic references, both on the Twitter user location field and their biography.

Ordun, Purushotham, and Raff (2020) studied techniques to assess the distinctiveness of topics, key terms and features, as well as the speed of dissemination of retweets over time. They used pattern matching and topic modeling with LDA on a set of 5.5M of tweets written in multiple languages, resulting in 16 topics for English and one for Spanish, Italian, French and Portuguese, respectively. They also applied Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville, 2018) to identify clusters of distinct topics, which discuss case spread, healthcare workers, and personal protective equipment issues.

Beyond Twitter, Amara, Taieb, and Aouicha (2020) have exploited 22K Facebook posts to track the evolution of COVID-19 related trends, with a multilingual dataset that covers seven languages (English, Arabic, Spanish, Italian, German, French and Japanese). They applied an end-to-end analytic process for discovering language-dependent topics covering the duration of the pre-crisis period and part of lockdown (from 1 January to 15 May of 2020). The experiments showed that the extracted topics corresponded to the chronological development of what has been happening, and the measures that were taken in various countries.

3 Methods

In what follows, we describe the methodology of our work, decomposed into four steps: (i) the collection of the corpus, (ii) the language identification and geolocation of the tweets, (iii) the preprocessing, and (iv) the topic modeling approach and its analysis, clustering tweets into topics and extracting representative keywords and sentences.

3.1 Collection of tweets

We first defined a set of keywords to download relevant tweets: *coronavirus*, *COVID-19*, *COVID19*, *2019-nCoV*, *2019nCoV*. Further, as of March 3th, 2020 we added more keys: *SARS-CoV-2*, *SARSCoV2*, *CoV-19*, *CoV19*, *COVD19*, *COVD 19*, *corona virus*, *corona outbreak*.

More particularly, we collected a multilingual corpus of 32.68M tweets, including Twitter posts from 1 January of 2019[1] to 20

[1] In order to have some preceding context, but ex-

April of 2020; from all over the world. We scraped the tweets using the GetOldTweets-python3 (GOT3) library.[2] The reason to use this tool was that it allowed to retrieve old tweets without time limitation. However, the tool did not permit us to filter the retrieval by language. Besides, the Twitter Official API cannot retrieve tweets more than a week ago with a free subscription mode.[3]

3.2 Language identification and geolocation

The next step is language identification to keep only the Spanish tweets. We used four tools for detecting languages, since with GOT3 we could not obtain the language attribute. Those four tools were: `polyglot`,[4] `langdetect`,[5] `langid.py`,[6] and `fastText`.[7] The language is assigned based on majority voting. In case of a tie, we consider the tweet to be Spanish, except if all tools predicted a different language.

In total, we identified 5.35M Spanish tweets. In this work, we try to restrict the analysis to the content generated in Spain. For this purpose, we proceeded to filter the tweets in Spanish using the location attribute of the user profile, and look for the name of Spanish cities with more than 50K inhabitants, province names, autonomous regions names, and also any location specified as simply 'Spain'.[8]

After the cleaning process, we obtained ∼1.85M tweets for our topic modeling analysis. It is fair to point out that there is a percentage of tweets with a risk of not being correctly filtered, since the same place name might exist in more than one Spanish speaking country (e.g., 'Guadalajara' for Spain vs. 'Guadalajara' for Mexico). This is a common limitation on Twitter analyses, when it comes to analyze geolocated tweets (see for instance (Vilares and Gómez-Rodríguez, 2018)).

3.3 Preprocessing

We first proceed to lowercase the tweets and remove retweets. We also delete the keywords that were used to collect the tweets (see again §3.1) and other Twitter reserved words such as 'rt', 'fav', 'vía', 'nofollow', 'twitter', 'href' or 'rel'. Moreover, we removed stopwords, non-words (i.e., words compounded with characters that are not alphabet letters), URLs, numbers and punctuation marks. To do this, we used spaCy[9] to tokenize the words, and the Spanish and English stopwords lists from three libraries: NLTK,[10] stop-words,[11] and stopwordsiso.[12] Besides, in order to remove extra noise and cluster more clean topics, we only kept content words (i.e., nouns, verbs, adjectives, and adverbs).

Finally, to reduce word sparsity we used a custom lemmatizer[13] for Spanish, which applies a rule-based lemmatization with spaCy, and relies on Wiktionary,[14] which is a collaborative free-content multilingual dictionary. After the lemmatization step, the tweets whose length is less than three characters were removed. As traditional topic modeling approaches such as LDA, based on bag-of-words, suffer if many outliers are present (which happens in NLP due to the Zipf's law), we ignore terms that have a corpus frequency strictly less than three.

3.4 Topic modeling

For a more clear and comprehensive topic modeling analysis, we cluster the tweets in four weeks per month, except for the year 2019 (for which we collect the few tweets discussing coronavirus topics at that time), and the month of January 2020, which covers the first fortnight and not a week.

More particularly, we cluster the time of analysis into three phases. First, a pre-crisis phase, which includes tweets up to 24 January of 2020; when there was still few cases

pecting just to be able to retrieve a small number of tweets.

[2] `https://github.com/Jefferson-Henrique/GetOldTweets-python`

[3] However, we noted that GOT3 as of 18 September 2020 has been suspended due to the new Twitter policies on tweet payload

[4] `https://polyglot.readthedocs.io/en/latest/Detection.html`

[5] `https://pypi.org/project/langdetect/`

[6] `https://pypi.org/project/langid/`

[7] `https://fasttext.cc/docs/en/language-identification.html`. We used the large model

[8] We obtained the list of place names from the Instituto Nacional de Estadística (INE) `https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525`

[9] `https://spacy.io/usage/v2-2` and the es_core_news_md language model

[10] `http://www.nltk.org/nltk_data/`

[11] `https://pypi.org/project/stop-words/`

[12] `https://pypi.org/project/stopwordsiso/`

[13] `https://github.com/pablodms/spacy-spanish-lemmatizer`

[14] `https://www.wiktionary.org/`

reported outside China. Second, we consider the outbreak phase, that we will consider to range from 25 January to 14 March of 2020; when the disease started to widely spread across Europe and the rest of the world, but Spain still was not under confinement. This is the period of time where the pandemic information, epidemic back then, was reported but was still not formally considered an alarm by the Spain government. Third, we cover about a month of the official lockdown period of the first wave (from 14 March to 20 April of 2020), when the Spanish government approved a strict social confinement.

As introduced previously, for topic modeling, we will be using *Latent Dirichlet Allocation* (LDA)[15] with collapsed Gibbs sampling inference (Griffiths and Steyvers, 2004); which processes raw text data in an unsupervised fashion to cluster documents that discuss the same topic. We chose LDA because it is standard and has probed robust for many tasks (see also §2.2 and §2.3). For each phase, we will mostly group tweets into weeks,[16] and for each week we will be extracting 10 topics. On the one hand, our goal was to facilitate the comprehension and interpretability. On the other hand, it is worth to note that selecting too few topics would make the clusters very generic and unspecific, while choosing too many could make them too sparse, not representative, and hard to analyze qualitative (Steinskog, Therkelsen, and Gambäck, 2017). Yet, we explored what would be in theory an optimal number of topics for different weeks using three methods: (i) the KL divergence (Arun et al., 2010), (ii) the pairwise cosine distance (Cao et al., 2009), (iii) and the loglikelihood. In all cases the results returned that the ideal number was between 5 and 20 in most of cases.

LDA setup We sampled up to 1500 epochs, and we kept the rest of parameters to the default value in the LDA library we used, i.e., $\alpha : 0.1$, $\eta : 0.01$, where the first corresponds to the Dirichlet parameter for the distribution over topics, and the second to the Dirichlet parameter for the distribution over words.

[15] In particular, we rely on the `https://github.com/lda-project/lda` implementation

[16] As introduced before, we use week here in an informal sense, referring to periods of time of 7 days, but not necessarily from Monday to Sunday.

3.5 Extracting top topic keywords and sentences

To extract the most representative keywords for each topic, we considered both generative (GS) (Equation 1) and discriminative (DS) (Equation 2) approaches:

$$\mathrm{GS(w,z)} = P(w|z) \quad (1)$$

$$\mathrm{DS}(w, z) = P(w|z)/[\max_{z' \neq z} P(w|z')] \quad (2)$$

where w represents a given word and z the topic at hand. In essence, the generative score allows to extract the words that are most representative for each topic independently, in a way that a given word could be relevant for one or more topics, potentially making such topics harder to differentiate among them. On the contrary, the discriminative score allows to represent a topic by a set of keywords that are very representative for such topic, but have little relevance for the remaining ones.

Although the top keywords for each topic are useful, they might provide a limited view of what is actually being discussed. To counteract this, we also defined a generative (Equation 3) and discriminative (Equation 4) routes to extract the most representative sentences (tweets) for each topic, ideally being able to determine the topic by simply reading a few documents. The motivation to define these two different routes is the same than the one we made to extract the top keywords.

$$\mathrm{GS}_{\mathrm{sent}}(s, z) = \sum_{w \in s} \mathrm{GS}(w, z)/\mathrm{Length}(s) \quad (3)$$

$$\mathrm{DS}_{\mathrm{sent}}(s, z) = \sum_{w \in s} \mathrm{DS}(w, z)/\mathrm{Length}(s) \quad (4)$$

where s is the input document, for which we consider its length, in order not to only select the longest documents; although in the case of Twitter this is less of an issue than in other topic modeling approaches that must deal with actual long documents.

The full code is available.[17]

Limitations Sociolinguistic studies that collect data from social media such as Twitter can suffer from biases that can be hard to measure, identify or correct. For instance, it is well-known that a small percentage of

[17] `https://github.com/mmaguero/twitter-analysis`

Topic	Discriminative Keywords	Generative Keywords
'W1' (from January to December of 2019)		
2	respiratorio, enfermedad, gripe	respiratorio, gripe, enfermedad
Magnifica guia para diferenciar los sintomas que causa la gripe y otros virus respiratorios. Junto con la gripe siguen circulando rinovirus, virus respiratorio sincitial y coronavirus, entre otros. <URL>		
1	enfermedad, gripe, respiratorio	enfermedad, respiratorio, gripe
@user informa de 27 casos de neumonia atipica, probablemente virica, en Wuhan (Hubei, China) en fecha 31/12/2019. El SARS (coronavirus) se inicio asi en 2003. Habra que seguir evolucion y esperar el diagnostico. <URL>		
W2-3 (from 1 to 16 January of 2020)		
8	alerta, hospital, poner, red, oms, china, mundial, mundo	china, oms, alerta, hospital, poner, mundial, mundo, red
UN NUEVO CORONAVIRUS PONE EN ALERTA A CHINA <URL>vía @user		
5	confirmar, japón, chino, infección, caso, china, animal, aparición	caso, confirmar, japón, china, infección, chino, ciudad, identificar
Japón confirma el primer caso de coronavirus vía @user <URL>		
W4 (from 17 to 24 January of 2020)		
9	emergencia, declaración, declarar, organización, reunión, convocar, decisión, determinar	oms, emergencia, internacional, declarar, mundial, alerta, salud, china
La OMS no declaró la emergencia por el coronavirus <URL>		
1	millón, cuarentena, habitante, frenar, ampliar, pekín, transporte, aislar	china, ciudad, wuhan, millón, cuarentena, persona, cerrar, brote
Más de once millones de chinos, en cuarentena por el coronavirus <URL>		

Table 1: Some representative topics for the weeks corresponding to the **pre-crisis** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

Twitter users generate the majority of content (Wojcik and Hughes, 2019). In this line, we believe that many of the collected tweets have its origin in newspapers and journalists accounts, that condition how other users tweet about this topic on Twitter, and therefore the detected topics can be heavily dependent on how national media decide to spread the news. Yet, this is the natural behaviour of this network, and in this particular work we decided not to control for this variable.

4 Results

We consider sixteen sets of tweets (mostly grouped in a weekly basis), extracting the ten most representative topics for each one according to LDA. To refer the topics, we will represent them with the top eight keywords and the most salient tweets. For clarity, and due to the large amount of weeks and topics, we will just illustrate and analyze some relevant topics extracted by our approach for different weeks, and try not to repeat common topics that span through the whole period. Usernames and urls are cut due to anonymity and space reasons, respectively.

4.1 Pre-crisis time

During this pre-crisis time, it is possible to see how the model captures that the COVID-19 was still not a concern for the Spanish society, which perceived the disease as an external problem, as reflected in many of the extracted topics. For clarity, Table 1 illustrates some relevant topics with top keywords and tweets, but we briefly discuss the content of the table below. To assess the relevance of the topics, we will be matching those against news from the newspapers that were published at the time in different Spanish media.

'W1' (from January to December of 2019) For the year 2019, we only could extract a total of seven topics, since the corresponding subset of tweets related to COVID-19 or coronavirus was still tiny (a total of 43 tweets after preprocessing). Still, we believe the results are interesting, since we observed that at this time most of Spanish tweets dealing with coronavirus still had to do with veterinarian diseases or even the zoonosis of coronavirus (i.e., how it is transmitted between animals and humans through the air), Yet, we found a few relevant tweets about COVID-19 that started to show up. We illustrate this as part of Table 1.

W2-3 (from 1 to 16 January of 2020) This time can be considered as the start of the emergency (Agencia EFE, 2020). In this line, we observed how our model started to identify this situation as well, clustering tweets about the World Health Organization (WHO) alerts to hospitals about symptoms, procedures, etc., and also about the increase

in the number of cases in China.

W4 (from 17 to 24 January 2020) The crisis started to expand and from our model we see how the topics differ from previous weeks (see the third group of rows in Table 1). For instance, it shows how China started to apply restrictions in many locations of its territory (e.g., Wuhan) (El Boletín, 2020).

4.2 Outbreak time

In this phase, we see how the LDA approach reflects emergency declarations, the first cancellations of massive events in Spain, as well as the first suspicious cases; causing in consequence an increase of the concern among the Spanish society, which started to look and ask for sanitary products. This is also the phase where the approach captures a transition from international to national concerns. We will breakdown this more in detail in the next paragraphs, matching again the topics against news from the newspapers to qualitatively verify the quality of the extracted topics. Table 2 illustrates such topics with the top keywords and tweets from the model.

W5 (from 25 to 31 January of 2020) During this week, the approach kept identifying online discussions about the WHO emergency declarations, considering COVID-19 as a global coronavirus threat (Pérez, 2020). Also, the approach extracted topics related to international restrictions, such as the airplane company Iberia suspending flights to Shanghai (CatalunyaPress.es, 2020), at the same time that Russia closed its frontiers with China (Ellyatt, 2020).

W6 (from 1 to 7 February of 2020) Following the trend of announcing emergency declarations, the model started to identify international issues, such as the infection and posterior death of Li Wenliang (BBC News, 2020), a Chinese doctor that alerted about the first cases of COVID-19 in December 2019, but also national ones; such as the confirmation of the first case of coronavirus in Spain, in the Canary Island of La Gomera (Linde, 2020). This matches the time where the number of cases seemed to start to spread (still slowly) all around the world.

W7 (from 8 to 14 February of 2020) During this week, the coronavirus started to have an important economic effect in Spain, which is reflected by the model, discovering topics that showed how users discussed the potential (finally confirmed during this week too) cancellation of the 2020 Mobile World Congress (MWC 2020), which usually takes place in Barcelona (Pardeiro, 2020). On the healthcare side, additional (few) cases started to be reported in Spain, such as in Mallorca, where it was reported the second Spanish case of COVID-19 (Bohórquez and Güell, 2020). During this and next weeks, we started to observe how there is a slow transition from international to national topics.

W8 (from 15 to 21 February of 2020) During this week, the topics where in line with those discussed in the previous weeks, such as the cancellation of the MWC 2020 (see Table 2) and its repercussion. This 'last-long' topics made sense at the time, since the cancellation of the MWC 2020 was the first massive event cancelled in Spain, with important economic consequences. Other international issues such as the sustained increase of cases in China or in the cruise ship Diamond Princess (Almoguera, 2020) seemed to occupy Twitter users during this time, too.

W9 (from 22 to 29 February of 2020) These are the final days before the lockdown period, and in retrospective, it is easy to see how some of the topics extracted reflected the immediate seriousness of the situation. We see how the model captures that the WHO advised to the public (World Health Organization (WHO), 2020a) to wash hands frequently. It is interesting to see in Table 2 how 'farmacia' (pharmacy) appears together with 'gel' (gel), 'lavarse' (to wash), 'mano' (hand) and 'alcohol' (alcohol), 'agotar' (to run out of) among the top keywords for the corresponding topic. In this context, it is well-known that these products were scarce in pharmacies and stores, and actually this problem lasted for long during the lockdown period. Also, related to the immediate seriousness of the situation, the model captured how despite of not being confined, the world economy started to suffer with the stocks set for the worst week since 2008 (Sano, 2020).

W10 (from 1 to 8 March of 2020) Just before the lockdown, we observe how among the topics extracted there are topics that we see everyday in the current pandemic life. For instance, as shown in Table 2, we kept seeing the importance of washing hands and keep a good hygiene with the use of soap (World Health Organization (WHO), 2020a). Also,

Topic	Discriminative Keywords	Generative Keywords
W5 (from 25 to 31 January of 2020)		
9	oms, emergencia, declarar, declaración, sanitaria, organización, comité, convocar	oms, internacional, emergencia, salud, declarar, alerta, mundial, china
Declara OMS emergencia por coronavirus - Vía @user <URL>		
1	vuelo, cerrar, suspender, frontera, kong, hong, rusia, aerolínea	china, vuelo, cerrar, suspender, brote, frontera, evitar, kong
Iberia suspende los vuelos a Shanghái por el coronavirus <URL>...		
W6 (from 1 to 7 February of 2020)		
4	alertar, acusar, news, silenciar, intentar, bbc, difundir, confusión	médico, china, chino, morir, alertar, wuhan, muerte, wenliang
Por favor lean. Porque esto no lo va a contar ningún medio que alerte sobre el coronavirus . <URL>...		
1	gomera, alemán, ingresado, contacto, jalisco, vitoria, ecuador, isla	caso, españa, gomera, paciente, sospechoso, hospital, salud, síntoma
En España ya tenemos un caso de coronavirus ,en La Gomera ,un alemán.		
W7 (from 8 to 14 February of 2020)		
6	mallorca, negativo, británico, ingresado, palma, princess, diamond, gomera	caso, mallorca, españa, crucero, paciente, confirmar, sospechoso, salud
Confirman un caso de coronavirus en Palma de Mallorca <URL>... <URL>		
3	sony, amazon, gsma, bajas, lg, nvidia, ericsson, intel	mobile, congress, barcelona, mwc, empresa, cancela, cancelar, sony
Tras las bajas de LG, Ericsson, NVidia, Amazon y Sony #coronavirus #MWC2020 <URL>...		
W8 (from 15 to 21 February of 2020)		
5	crucero, diamond, princess, pasajero, colombiano, camboya, evacuado, ucrania	crucero, cuarentena, japón, caso, diamond, princess, pasajero, wuhan
NUEVOS CASOS DE CORONAVIRUS EN CRUCERO DIAMOND <URL>... <URL>		
6	mobile, barcelona, cancelación, cancelar, maratón, evento, mwc, congress	mobile, china, tokio, barcelona, cancelar, cancelación, maratón, guerra
Suspenden el Mobile World Congress de Barcelona por el coronavirus <URL>... <URL>		
W9 (from 22 to 29 February of 2020)		
6	mano, farmacia, lavarse, gel, desinfectante, alcohol, agotar, carne	mascarillas, mano, gente, mascarilla, evitar, comprar, miedo, hospital
Cómo prevenir el #coronavirus . Lávate las manos, lávate las manos, lávate las manos..... lávate las manos. <URL>... <URL>		
1	bolsa, economía, mercado, caída, ibex, pérdida, crecimiento, wall	china, bolsa, economía, mercado, crisis, mundial, impacto, económico
'Esto es mercado. Esto me pone' @user #bolsa #COVID19 <URL>...		
W10 (from 1 to 8 March of 2020)		
7	mano, metro, lavarse, higiene, agua, gel, jabón, lavar	mano, evitar, medido, contagio, mascarillas, covid, persona, tomar
me voy a lavar las manos que no quiero el coronavirus		
4	patología, contagioso, anciano, letalidad, diferencia, estacional, comparación, hambre	gripe, persona, año, morir, mortalidad, gente, matar, enfermedad
Se llama Virus Corona Patologías Previas		

Table 2: Some representative topics for the weeks corresponding to the **outbreak** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

'metro' (underground) is a top keyword of such topic, since at that time there was a discussion about the chances of getting infected (e.g., in the public transport) (CNN, 2020). In a different topic, we see what it seems to be a discussion comparing the flu and covid, and how they affect to the population, which was a popular comparison at the time.

4.3 Lockdown time

During the lockdown phase (until April), we can observe in Table 3 how the topics discussed mostly focused on the worst consequences of the pandemic, such as the big economic crisis, the large number of deaths per day, and also some collective actions such as thanking the healthcare workers. Again, we give a brief explanation below these lines, and match the topics against news in the media.

W11 (from 9 to 16 March of 2020) Here we consider the week where the Spanish society stopped to have free movement. More particularly, the government approved strict social confinement on 14 March of 2020 (Cué, 2020). Besides, the model found topics about the acknowledgement to the healthcare workers and the solidarity applause (La Razón, 2020), which was very popular in Spain dur-

Topic	Discriminative Keywords	Generative Keywords
W11 (from 9 to 16 March of 2020)		
3	aplauso, frenalacurva, aplausosanitario, cuarentenaya, yoelijoserresponsable, felizlunes, arena, agradecimiento	covid, yomequedoencasa, casa, quedateencasa, cuarentena, coronavirusesp, cuarentenacoronavirus, responsabilidad
Aplausos para que suenen más que los truenos que hoy hay en Madrid. Hoy mis aplausos para todos. Para \Saldremos de esta / #COVID19 <URL>		
8	ocasionado, aprobar, pymes, paliar, erte, fiscal, boe, hipoteca	covid, medido, crisis, gobierno, alarma, situación, sanitaria, empresa
#RealDecreto 463/2020 #estadodealarma #COVID19 <URL>#pymes #Autonomo #Cordoba @user <URL>		
W12 (from 17 to 24 March of 2020)		
6	respirador, fabricar, ifema, impresora, envío, coronavirius, epis, todosobremovil	covid, hospital, sanitario, mascarillas, madrid, personal, estevirusloparamosunidos, quedateencasa
#ElonMusk puede que empiece a fabricar respiradores #COVID19 <URL>		
1	higiene, jabón, distanciamiento, acatar, lavado, fanb, geacam, comerciales	covid, medido, evitar, contagio, prevención, propagación, salud, tomar
Entre más higiene se tenga, mayor es la protección ante los patógenos como el #COVID19 <URL>...		
W13 (from 25 to 31 March of 2020)		
1	erte, pago, despido, prestación, contrato, fiscal, ertes, alquiler	crisis, medido, gobierno, empresa, económico, trabajador, autónomo, sanitaria
Información para los afectados por ERTE debido al COVID19 . #ERTE #Coronavirus <URL ><URL>		
3	civil, guardia, desinfección, desinfectar, higiene, cumplimiento, jabón, estación	medido, persona, evitar, contagio, salud, seguridad, prevención, casa
Unos 400 guardias civiles con coronavirus en #CLM , según la @user @user -. Vía @user <URL>... <URL>		
W14 (from 1 to 7 April of 2020)		
1	animal, respiratorio, tigre, gato, mascota, zoo, bronx, contaminación	persona, paciente, enfermedad, síntoma, casa, contagio, evitar, matar
Si los tigres se contagian de coronavirus , ¡ojito los que tenéis gato!		
8	confirmado, cifra, elevar, defunción, diarios, ascender, activos, descender	caso, fallecido, españa, muerte, muerto, dato, número, país
637 muertes por coronavirus en un día, la cifra más baja en 13 días <URL>		
W15 (from 8 to 14 April of 2020)		
2	cifra, curado, reino, récord, ascender, contabilizar, diagnosticado, acumular	caso, fallecido, españa, muerto, muerte, dato, número, persona
Las 510 muertes por COVID-19 en un día, la cifra más baja desde el 23 de marzo <URL>		
5	johnson, intensivo, boris, testimonio, alta, universitario, clmpressdigital, sosprisiones	hospital, médico, paciente, sanitario, madrid, personal, persona, profesional
Coronavirus : Boris Johnson fue dado de alta. <URL>		
W16 (from 15 to 20 April of 2020)		
5	luis, sepúlveda, escritor, homenaje, chileno, fútbol, club, dep	año, morir, hospital, luis, fallecer, quedateencasa, yomequedoencasa, historia
Luis Sepúlveda muere por coronavirus <URL>... <URL>		
2	distanciamiento, prohibidorendirse, enestafamilianadieluchasolo, yonosoyungastosuperfluo, bicicleta, espandemia, comunidadvalenciana, saltarse	confinamiento, quedateencasa, yomequedoencasa, cuarentena, medido, casa, evitar, alarma
¿El distanciamiento social podría ir incluso más allá de 2021? #COVID19 #coronavirus <URL>		

Table 3: Some representative topics for the weeks corresponding to the **lockdown** period of the COVID-19 pandemic in Spain. For each example topic, we include the top representative sentence according to its discriminative score.

ing the lockdown period. In a related line, topics like this one also captured the feeling of the importance of staying at home to prevent becoming infected and reduce the workload of these workers.

W12 (from 17 to 24 March of 2020) For this week, the model extracted topics discussing the personal hygiene measures to combat the COVID-19. The topics also reflect the lack of equipment in the hospitals, which was a problem at the beginning of the pandemic. More particularly, the model was able to identify as a topic the lack of ventilators in Spain, and also the rest of the world, as reflected by the most salient discriminative tweet. This matches the news at the time, which discussed the use of 3D printers to provide such ventilators (Polo, 2020), or hacking some objects to adapt them for medical use (Cristian Fracassi, 2020).

W13 (from 25 to 31 March of 2020) This week covers the last days of March 2020.

Due to the strict confinement, topics concerning job losses and the measures taken by the government to counteract the situation (e.g. the so-called ERTEs) started to arise (RTVE.es, 2020; Gestiona.es, 2020). Among the rest of the topics of this week, we also would like to remark the massive infection of public workers, such as the Guardia Civil officers in Castilla La-Mancha (EFE/CMM, 2020). The infection of public workers during this time of the pandemic was also widely discussed in the news (Requeijo, 2020).

W14 (from 1 to 7 April of 2020) On the national side, some topics reflected the number of casualties per day. More particularly, the beginning of April corresponded to the peak of the first wave, and the beginning of the decreasing trend in the number of infections and deaths per day (Justo, 2020). A bit on a different line, we found topics discussing more diverse aspects of COVID-19, such as the infection in the Zoo of Bronx (New York, USA) of tigers and lions (M.R.M., 2020).

W15 (from 8 to 14 April of 2020) Here, we would like to remark a topic related to an international breaking news, and more particularly about Boris Johnson (the UK Prime Minister) being infected by the coronavirus, together with his evolution, when he even entered the ICU (La Vanguardia, 2020). On the national side, the models kept detecting topics related to the number of deaths in Spain, which was still high and dynamic during that time, but reached some local minima these days (Soteras, 2020).

W16 (from 15 to 20 April of 2020) For the last days of our study, the model found relevant topics too, such as the death of the Chilean writer Luis Sepúlveda (Safont Plumed, 2020) due to COVID-19, or topics related to the need of keeping social distancing, maybe even for months (elEconomista.es, 2020), as reflected by some of the most representative tweets.

4.4 Quantitative evaluation

We performed a small human evaluation to quantitatively estimate the quality of the extracted topics. We took 20 topics randomly from all periods. Then, two annotators were in charge of: (i) determining if given the top 8 keywords and 3 top sentences made possible to infer a topic, (ii) determining if for each top topic word (according to the discriminative score) they belonged to the inferred topic, and (iii) the same as in (ii), but for the 3 most representative sentences. We calculated the percentage of times both annotators positively labelled a sample, obtaining scores of 80%, 56.88% and 71.66% for (i), (ii), and (iii), respectively. In addition, we calculated (ii) but taking into account only the first 3 top keywords of the topic, yielding a score of 75% of positive samples.

5 Conclusion

This paper used a topic modeling approach to shed some light about the topics discussed in Spain during the early stages of the COVID-19 pandemic, including a period of pre-crisis, the outbreak of the disease, and the beginning of the confinement. We collected a large amount of tweets using keywords and cleaned them to keep only Spanish tweets that were written in Spain. After that, we used a Latent Dirichlet Allocation model that learned to cluster such tweets according to the topic they discuss. To represent the topics, we used generative and discriminative routes to extract the most salient keywords and sentences. To verify the quality of the extracted topics, we performed a qualitative analysis matching the topics against relevant news in the newspapers at the same period of time, and a small quantitative evaluation. Overall, the topics show that during the pre-crisis period, users focused on the international panorama than the local situation, while during the outbreak and lockdown phases they focused the most on the Spanish emergency, considering health and economic problems.

Acknowledgements

MMAT has been partially funded by Barcelona Supercomputing Center (BSC) through the Spanish Plan for advancement of Language Technologies 'Plan TL' and the Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA). DV is supported by MINECO (TIN2017-85160-C2-1-R), by Xunta de Galicia (ED431C 2020/11), by Centro de Investigación de Galicia 'CITIC' (European Regional Development Fund-Galicia 2014-2020 Program, ED431G 2019/01), and by a 2020 Leonardo Grant for Researchers and Cultural Creators from the BBVA Foundation.

References

Abd-Alrazaq, A., D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah. 2020. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016.

Afzal, Z., V. Yadav, O. Fedorova, V. Kandala, J. van de Loo, S. A. Akhondi, P. Coupet, and G. Tsatsaronis. 2020. CORA: A deep active learning covid-19 relevancy algorithm to identify core scientific articles. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Agencia EFE. 2020. La OMS pone en alerta a la red mundial de hospitales por un nuevo coronavirus en China. *www.efe.com*, January.

Almoguera, P. 2020. El coronavirus pone en jaque ahora a Japón y Corea del Sur. *El País*, February.

Amara, A., M. A. H. Taieb, and M. B. Aouicha. 2020. Multilingual topic modelling for tracking covid-19 trends based on facebook data analysis.

Andrzejewski, D. and D. Buttler. 2011. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 600–608.

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer.

Asgari-Chenaghlu, M., N. Nikzad-Khasmakhi, and S. Minaee. 2020. Covid-transformer: Detecting trending topics on twitter using universal sentence encoder. *arXiv preprint arXiv:2009.03947*.

Banda, J. M., R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalina, and G. Chowell. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration, August.

Barde, B. V. and A. M. Bainwad. 2017. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750.

BBC News. 2020. Li Wenliang: Coronavirus kills Chinese whistleblower doctor. *BBC News*, February.

Blei, D. M. and J. D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bohórquez, L. and O. Güell. 2020. El segundo caso de coronavirus en España es un británico que se contagió en los Alpes. *El País*, February.

Boon-Itt, S. 2020. A text-mining analysis of public perceptions and topic modeling during the covid-19 pandemic using twitter data. *JMIR public health and surveillance, JMIR Preprints. 30/06/2020:21978*.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang. 2009. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Carbonell Gironés, L. 2020. Geographical analysis of the opinion and influence of users on twitter during the coronavirus health crisis. Final project/degree, Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València.

CatalunyaPress.es. 2020. Iberia suspende los vuelos a Shanghái por el coronavirus.

Cer, D., Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder.

Chandrasekaran, R., V. Mehta, T. Valkunde, and E. Moustakas. 2020. Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. *Journal of Medical Internet Research*, 22(10):e22624.

Chen, E., K. Lerman, and E. Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

CNN. 2020. Medidas globales por el coronavirus: mantener distancia de un metro, cierre de escuelas y museos, evitar los besos y otras, March.

Cristian Fracassi. 2020. Charlotte valve, March.

Cué, C. E. 2020. El Gobierno informa de que es la única autoridad en toda España, limita los desplazamientos y cierra comercios, March.

Deerwester, S. 1988. Improving information retrieval with latent semantic indexing.

EFE/CMM. 2020. 400 guardias civiles de Castilla-La Mancha tienen Covid-19, según la AUGC.

El Boletín. 2020. China pone en cuarentena a más de 30 millones de personas por el coronavirus. January.

elEconomista.es. 2020. Las medidas de distanciamiento social podrían extenderse hasta 2022 de manera intermitente - elEconomista.es.

Ellyatt, H. 2020. Russia closes border with China to prevent spread of the coronavirus, January.

Gao, Y., Y. Xu, and Y. Li. 2014. Pattern-based topics for document modelling in information filtering. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1629–1642.

Gestiona.es. 2020. Información para los afectados por ERTE debido al COVID19, March.

Griffiths, T. L. and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Grover, P., A. K. Kar, Y. K. Dwivedi, and M. Janssen. 2019. Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438–460.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hutto, C. and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, volume 81, page 82.

Jagarlamudi, J., H. Daumé III, and R. Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Justo, D. 2020. España sigue la tendencia a la baja: 4.273 nuevos contagios por coronavirus y 637 muertes, April.

Kerchner, D. and L. Wrubel. 2020. Coronavirus Tweet Ids.

Kleinberg, B., I. van der Vegt, and M. Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

La Razón. 2020. Emotivo reconocimiento a los sanitarios en forma de aplausos desde los balcones, March.

La Vanguardia. 2020. Boris Johnson recibe el alta y continuará recuperándose de la Covid-19 en su casa, April.

Linde, P. 2020. Sanidad confirma en La Gomera el primer caso de coronavirus en España. *El País*, February.

Loria, S. 2020. textblob documentation. *Release 0.16*, 2.

McInnes, L., J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February.

Mohammad, S. M. and P. D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Moody, C. E. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec.

M.R.M. 2020. Un tigre del zoo de Nueva York tiene coronavirus, April.

Neubig, G., Y. Matsubayashi, M. Hagiwara, and K. Murakami. 2011. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973.

Ordun, C., S. Purushotham, and E. Raff. 2020. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.

Pardeiro, M. 2020. El fracaso político del MWC: "No se va a suspender". "No cuelga de un hilo".

Pham, P., P. Do, and C. D. Ta. 2018. W-pathsim: novel approach of weighted similarity measure in content-based heterogeneous information networks by applying lda topic modeling. In *Asian conference on intelligent information and database systems*, pages 539–549. Springer.

Polo, J. 2020. Coronavirus: La Zona Franca fabricará 100 respiradores diarios con impresoras 3D, March.

Pérez, B. 2020. La OMS rectifica y declara la emergencia global por el coronavirus, January.

Requeijo, A. 2020. La Policía y la Guardia Civil suman ya más de 400 positivos por coronavirus, March.

RTVE.es. 2020. Los ERTE por la crisis del coronavirus suman más de 240.000, March.

Safont Plumed, J. 2020. Muere el escritor chileno Luis Sepúlveda, a causa del coronavirus.

Salton, G., A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.

Sano, H. 2020. GLOBAL MARKETS-World stocks set for worst week since 2008 as virus fears grip markets. *Reuters*, February.

Soteras, A. 2020. COVID-19: 510 muertes en un día, la cifra más baja desde el 23 de marzo.

Steinskog, A., J. Therkelsen, and B. Gambäck. 2017. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st nordic conference on computational linguistics*, pages 77–86.

Tuarob, S., L. C. Pouchard, and C. L. Giles. 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 239–248.

Verspoor, K., K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea, and B. Wallace, editors. 2020a. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Verspoor, K., K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, and B. Wallace, editors. 2020b. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Vilares, D. and C. Gómez-Rodríguez. 2018. Grounding the semantics of part-of-day nouns worldwide using twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 123–128.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Wojcik, S. and A. Hughes. 2019. Sizing up twitter users. *PEW research center*, 24.

World Health Organization (WHO). 2020a. Advice for the public on COVID-19 – World Health Organization.

World Health Organization (WHO). 2020b. WHO statement regarding cluster of

pneumonia cases in Wuhan, China. January. Accessed: 2020-08-28.

Yijun, G. and X. Tian. 2014. Study on keyword extraction with lda and textrank combination. *Data Analysis and Knowledge Discovery*, 30(7):41–47.

Yin, H., S. Yang, and J. Li. 2020. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. *arXiv preprint arXiv:2007.02304*.

Yu, J., Y. Lu, and J. Muñoz-Justicia. 2020. Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo. *International Journal of Environmental Research and Public Health*, 17(15):5414.

Zhou, S., K. Li, and Y. Liu. 2009. Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, 2(4):398–409.

A.4 Deep learning and multilingual sentiment analysis on social media data: An overview

- Agüero-Torales, M. M., Abreu Salas, J. I., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, 107373. doi: https://doi.org/10.1016/j.asoc.2021.107373.
  - Status: Published.
  - Impact Factor (JCR 2020): 6,725.
  - Subject Category: Computer Science, Artificial Intelligence. Ranking 24/139 (Q1).
  - Subject Category: Computer Science, Interdisciplinary Applications. Ranking 11/111 (Q1).
  - Scopus CiteScore (2020): 11,2.
  - Subject Category: Computer Science, Software. Ranking 32/389.

A.4.1 Summary

This work was submitted to the 'Special Issue on Soft Computing for Recommender Systems and Sentiment Analysis'[6] of the journal 'Applied Soft Computing',[7] the submitted review's goal is to provide the community with an in-deep review of the works that had leveraged advances on deep learning to tackled the problem of multilingual sentiment analysis in social media.

There have been recent efforts to systematize deep learning approaches for sentiment analysis, for example, Agarwal et al. (2020) [127]. Though they also analyzed some samples for the multilanguage case, there was a large corpus of research and ideas that were not covered. Moreover, our work provided a high-level perspective of the field, identifying common ideas and problems that have been addressed to perform multilingual SA.

The presented literature review discussed the various Deep Learning approaches used to deal with multilingual settings, such as multi-language (those trained in several languages at once), cross-lingual (trained in one language and transferred to another), and code-switching (trained in one or several languages present at code-switching), with a focus on the independence of languages and linguistic features (in a language-agnostic way).

Thus, our work contributed to the research community with a review that:

i. Provides insights for the different approaches or setups for multilingual sentiment analysis, i.e. purely multilingual, cross-lingual, and code-switching. This distinction is useful because each approach follows a different hypothesis that in turn, influences the strategies developed to tackle the problem. This is more noticeable for the aspect-based level and the code-switching setup.

ii. Identifies the underlying idea behind solutions unrelated at first glance. This helps to better compare works and results, whether an idea has been widely adopted or not, as well as its limitations and the best and worst scenarios. Also, to ease the identification of new approaches and contributions.

iii. A reader-friendly summary of the languages, techniques, social media, and corpora that have been covered. This allows a quick yet comprehensive view of the field, valuable for all the public interested in this domain. It also helps to identify possible new research niches.

[6] https://www.sciencedirect.com/journal/applied-soft-computing/special-issue/10JLLL13821
[7] https://www.journals.elsevier.com/applied-soft-computing

Deep Learning and Multilingual Sentiment Analysis on Social Media Data: An Overview

Marvin M. Agüero-Torales[a,*], José I. Abreu Salas[b], Antonio G. López-Herrera[a]

[a] *Dept. of Computer Science and Artificial Intelligence, University of Granada, Calle Daniel Saucedo Aranda, s/n, 18071, Granada, Spain*
[b] *University Institute for Computing Research, University of Alicante, Carretera de San Vicente del Raspeig s/n, Alicante, Valencia, Spain.*

Abstract

Twenty-four studies on twenty-three distinct languages and eleven social media illustrate the steady interest in deep learning approaches for multilingual sentiment analysis of social media. We improve over previous reviews with wider coverage from 2017 to 2020 as well as a study focused on the underlying ideas and commonalities behind the different solutions to achieve multilingual sentiment analysis. Interesting findings of our research are (i) the shift of research interest to cross-lingual and code-switching approaches, (ii) the apparent stagnation of the less complex architectures derived from a backbone featuring an embedding layer, a feature extractor based on a single CNN or LSTM and a classifier, (iii) the lack of approaches tackling multilingual aspect-based sentiment analysis through deep learning, and, surprisingly, (iv) the lack of more complex architectures such as the transformers-based, despite results suggest the more difficult tasks requires more elaborated architectures.

Keywords: Sentiment Analysis, Multilingual, Cross-lingual, Code-switching, Deep Learning, Natural Language Processing (NLP), Social Media

*Corresponding author
Email addresses: maguero@correo.ugr.es (Marvin M. Agüero-Torales), ji.abreu@ua.es (José I. Abreu Salas), tel. +34-958-248557; fax: +34-958-243317 (Antonio G. López-Herrera), lopez-herrera@decsai.ugr.es (Antonio G. López-Herrera)

1. Introduction

Sentiment Analysis (SA) allows us to automatically evaluate the opinion of peoples toward products, services, and other entities. This knowledge can help to make better decisions looking to improve key performance indicators. Besides, the massive adoption of social media such as Facebook and Twitter, platforms for e-commerce and services like Amazon, and even review-specialized sites such as Rotten Tomatoes, unleashed a vast amount of content to be analyzed. This data is naturally multilingual and multicultural thus, an analysis based on a single language may carry the risk of not capturing the overall insights [1]. Moreover, important challenges can prevent fully leverage this data. Except for a few cases, e.g., English, most languages lack well-maintained resources widely used for SA such as annotated corpus and lexicons. Second, it could be not straightforward to adapt the same SA model to different languages, for example, due to variations in word order or usage, or the noise introduced by machine translation. Also, we have code-switching content, where users express their opinions using a mixture of languages in the same sentence.

Multilingual Sentiment Analysis (MSA) is an attempt to address those issues through several strategies. For example, taking advantage of resource-rich languages to perform SA in a resource-poor language as characteristic in cross-lingual sentiment analysis. Also, developing language-independent models capable to handle SA in different languages or a code-switching setup.

There is a wide spectrum of approaches for SA, for example, [2, 3, 4, 5], which can be relied on supervised but also in unsupervised methods that exploit sentiment lexicons, grammatical analysis, and syntactic patterns. In section 2.1 and 2.3 we include a panoramic of the different formulations of this task as well as the evolution of SA and MSA. More recently, deep learning (DL) approaches have become a trend leading to state-of-the-art results, with authors such as [6, 7, 8] exploring Convolutional Neural Networks, Adversarial Networks, and Recurrent Neural Networks among other models. In section 2.2 we resume some of the advancements of deep learning for SA as an introduction for the main

topic of this work, the applications of deep learning in multilingual sentiment analysis in social media.

Using the methodology detailed in section 3 as a guideline, we curated and reviewed 24 relevant research papers. We categorized them as regard the main idea in multilingual, cross-lingual or code-switching approaches, covered in 4.1, 4.2 and 4.3 respectively. For each one, we discuss its distinctive contributions, the experimental setup, corpus, and main results. Also, section 4.4 includes a comparative that allows a quick and broad view of the advances in the domain. This analysis drew interesting conclusions such as the few works, to date, leveraging recent developments in contextual embedding. Other main findings and conclusions are covered in sections 5 and 6.

As sentiment analysis and deep learning approaches have been growing as an important research field, there have been early efforts [1, 9, 10, 11, 12, 13] to systematize the knowledge corpus in this domain, works extended lately by [14, 15, 16]. Recently, [17] examined the fundamentals of the multilingual case. However, more than seventeen works we identified introducing or exploring specific ideas for MSA have not been studied by the aforementioned works. Another of our main contributions is to drive the review by the underlying hypothesis of each work, not only analyzing them as regards the type of neural network they used. This is important since the same task can be tackled by very different ideas. Also, we focused the analysis on the current three major strategies for MSA: multilingual, cross-lingual, and code-switching. This high-level view of the domain can help to unveil interesting patterns more than the type of neural network implemented. For example, the use of adversarial training to learn language-agnostic features.

2. Preliminaries

In this section, we cover the fundamental concepts in Sentiment Analysis and Multilingual Sentiment Analysis. Also, some of the antecedents about the applications of Deep Learning to this problem.

2.1. Sentiment analysis on social media

Starting from Wiebe et al. [18] work in the late 90s, there has been a surge of interest in the different setups of SA. In general, it can be done at a document, sentence, or aspect level [5] and the classification in terms of positive, negative, or neutral, but also other more fine-grained scales such as a ranking from 1 to 5. This attention over SA is closely tied to Social Media in its key role in the rise of modern SA particularly with the works of Pang et al. [19], and Turney [20], in 2002. The first used machine learning (ML) classification techniques over movie review data outperforming human-produced baselines. The second achieved an average accuracy of 74% for his recommendations based on online reviews, which used Semantic Orientation (SO) applied to unsupervised classification. Later, Pang and Lee (2008) [2], focused on the fundamentals and basic applications of SA, with a list of resources such as lexicons or datasets.

A comprehensive review that shows the maturity of SA up to 2012 can be found in the book of Liu [4]. This work covers most of the topics, definitions, research problems (e.g., opinion spam detection), types of opinions (such as explicit and implicit opinions), and classification algorithms for SA. In 2013 Feldman [21] and Cambria et al. [22] wrote about the basic techniques, key tasks, and applications as well as the evolution of the field.

Another source to take the pulse of the continuous advances in the field has been the tasks related to SA in Twitter hosted by the International Workshop on Semantic Evaluation (SemEval) from 2013 to 2017 and in 2020 for code-switching text. From the latest results, we can corroborate a shift toward the application of deep learning with 20 out of 48 systems participating in SemEval 2017 [23]. In the next section, we overview some of the recent advances in DL applied to SA without considering the multilingual task.

2.2. Deep learning on sentiment analysis task

Deep Sentiment Analysis (DSA) relies on the great potentials of DL showed for NLP tasks. Here, we briefly commented on some examples to illustrate how DL has been leveraged within the SA.

Word embeddings are used for language modeling and feature learning. They are commonly used as an input of the DL models, being Word2Vec [24] and GloVe [25] two frequently used approaches. Also, there are contextualized embeddings such as ELMo [26], which represent better the polysemy of the words. Besides using pre-trained embeddings, they can be learned to encode some specific task semantics. In the context of SA, this approach has been explored in works such as [27] and [28].

Another trending field within DL is the attention mechanism, which allows the model to non-uniformly weigh the contribution of the context when computing the output. This is another choice that is being used frequently in SA, for example, to capture the interaction between aspects and their context as in [29], [30].

Also, there has been a great interest in novelty architectures for SA. One approach that has been received considerable attention when working at the document level, is the design of hierarchical models which learn a representation for sentences from its words, on top of this level, another model can learn representations for documents. Different alternatives such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) can be used at each level. Works in [31], [32] and [33] are some examples of this approach.

As the last additional example of the ideas that have been explored, not only for monolingual SA but also for MSA we mention the use of adversarial learning to produce a set of domain-independent features. This is the hypothesis of works such as [34] and [35] for cross-domain SA.

The concepts discussed in this section do not exhaust the applications of DL to SA, a more complete revision can be found in [36].

2.3. Multilingual sentiment analysis

In this section, we introduce the fundamentals of multilingual sentiment analysis as well as some of the earlier applications of DL to MSA. Initially, the applications of SA have been developed basically for one language, English in most cases, but the multilingual nature of Social Media has shifted the field to a

multilingual analysis. Also, advances in SA backed by DL have made it possible to include low resource languages and avoid the use of translation tools.

A frequent approach for MSA is called Cross-Language (also Cross-Lingual) Sentiment Classification [37] which relies on machine translation [38, 39]. For example, [40] reported an improvement over non-DL classifiers (SVM - Support Vector Machine) translating from other languages (Hindi, Marathi, Russian, Dutch, French, German, Portuguese, Spanish, and Italian) to English to use augmented word embeddings together a CNN model. However, the cross-language approach carries several issues and weaknesses, for example, notable discrepancies in the data distribution, potential cultural distances even in a perfect translation, hard and costly translation tasks for large corpora with issues as charges, availability, performance [37].

Another usual setup for SA is the code-switching one, also called code-mixing or code-mixed. In this case, the content to be analyzed is expressed alternating two or more languages even in a single sentence. One early approach that takes on this problem with DL is Wang et. al [41] for Chinese-English. They proposed a bilingual attention LSTM to perform SA in a corpus from Weibo.com capturing the informative words from both the bilingual and monolingual contexts. Code-mixing is frequent in social media sites such as Facebook and Twitter in countries with a large part of its population speaking more than one language, for example in India where several official and non-official languages are used. This is an active research area for Hindi-English. An early proposal [42] is based in MLP (Multi-Layer Perceptron) with word-level features for Hindi-English and Bengali-English Facebook posts. A more detailed summary of SA for Indian languages with a special focus on the code-mixed text can be found in [43].

There has been a lot of interest to systematize the advances in MSA. For example, [44] and [45], the latter reviewed the principal directions of research focusing on the development of resources and tools for multilingual subjectivity and SA and addressed both multilingual and cross-lingual methods. Singhal & Bhattacharyya (2016) [11] described some of the different approaches used in

SA research. Lo et. al (2017) [1] revised various of the main approaches and tools used for MSA at the time. They identified challenges and provided several recommendations with a framework for dealing with scarce resource languages. Also, in [13] and [46] we can find revisions of the field, however, they did not delve into the use of DL in MSA.

3. Methodology

Our research methodology comprises four steps to realize our main goal: to identify the underlying ideas and commonalities behind the different solutions to achieve multilingual sentiment analysis as well as to suggest future research directions.

(i) Define the research scope: the applications of deep learning to multilingual sentiment analysis on social media from January 2017 to December 2020. We choose this timeframe because the shift toward the application of DL-based sentiment analysis happened in 2017 as showed in [23].

(ii) Article search: our search terms were, (a) deep learning AND (b) sentiment analysis AND (c.1) multilingual OR multi-language OR multilanguage; (c.2) crosslingual OR cross-lingual OR cross-language OR crosslanguage; (c.3) code-mixed OR codemixed OR code-mixing OR codemixing OR code-switching OR codeswitching; (c.4) bilingual OR bi-lingual. Search queries were run in Scopus and the Web of Science.

(iii) Article verification: the search yielded 96 studies that were examined to ensure they satisfy the following criteria. (a) must handle explicitly the multilingualism either by (a.1) training with one or more languages and evaluating with a different one or others, (a.2) train with a multilingual corpus, i.e., the same model sees text in different languages during training regardless of this being at different steps. Thus, we excluded works that separately trained and evaluated the same architecture in different languages, i.e., created one model for each language trained only with data from the given language. Candidates for deletion were verified by the three authors.

We also revised the citations from the selected studies as well those referenced by previous reviews [1, 12, 14, 15, 16, 17, 46, 47, 48, 49] to identify possible candidates, applying the filters (a.1, a.2). In total, 24 publications were eligible for review.

(iv) Research analysis: for each of the selected papers we extracted data and information about (a) research characteristics, as authors, year of publication, languages covered, methodology, corpus characteristics; (b) sentiment level and categorization (binary, ternary, or fine-grained, e.g., rates $0-5$); (c) deep learning architectures and techniques, and (d) results and effectiveness of the proposal against baselines or state-of-the-art models. We also reviewed each work to identify the underlying idea to handle multilingualism, with a focus on the results that assessed the hypothesis.

4. Deep learning techniques for multilingual sentiment analysis on social media

In this section, we review a large corpus of research related to the applications of deep learning to multilingual sentiment analysis. Instead of driving the analysis by the type of architecture or techniques, we choose to organize the works by its sub-domain within MSA, i.e., multilingual, cross-lingual, or code-switching approaches. Inward each category, we proceeded chronologically to track the evolution of the field but separating the aspect-based studies since in general, they lead to very specific architectures. Also, we aim to provide a high-level perspective considering the underlying hypothesis of each work. Finally, the epigraph 4.4 aids the reader to take a glimpse of the domain as regards the models, baselines, corpus, core ideas, and languages covered. For clarity, and due to the variety of datasets and languages covered by each work, Table 1 illustrates the corpus and reports the number of tweets/sentences/documents used.

4.1. Multilingual approaches

This category groups a large set of works that aims to be language-agnostic to those seen during the training. Common goals are the design of systems capable to learn directly from multilingual unpaired content and providing predictions regardless of the source language. Across the analysis, we will use acronyms or abbreviations of common domain concepts without their definition for the sake of space.

4.1.1. Sentence-based studies

Training the same model for different languages is explored by[50]. They fit a multilayered CNN in two phases. First, they learn word embeddings from a large corpus of $300M$[1] unlabeled tweets in English, Italian, French, and German. The parameters are optimized further during the second stage, trying to infer weak labels inferred from emoticons. Finally, they fine-tuned the model using a corpus of annotated tweets. Experiments evaluated a model (FML-CNN) trained in all languages at once, a model fitted in a single language (SL-CNN), and other variations. The results showed that FML-CNN reaches slightly worse performance, about 2.45% lower F1 score, compared to SL-CNN (67.79% for Italian). However, experiments suggest that FML-CNN can handle better for code-mixed text.

Another hypothesis is to exploit character level embeddings to achieve language independence. In [37] and [51] authors describe language-agnostic translation-free architectures (Conv-Char-S, Conv-Char-R) for Twitter based on a CNN that can be trained in several languages at once. They evaluated their approach using tweets in English, Portuguese, Spanish and German from the corpus in [52] achieving an F1 score above 72.2% [51] for the multilanguage setup. The slightly worst results for some baselines such as LSTM-Emb [53] can be a trade-off since the models have ≈ 90 times fewer parameters and use ≈ 4 times less memory.

[1] M: Million.

The idea of multilanguage character embeddings is explored also by [54] but mapping each character to its UTF-8 integer code. The architecture (UniCNN) is similar to [37, 51], placing a CNN after the embedding layer, with a fully connected classifier at the top. They used a subset of the Twitter corpus in [52]. The UniCNN achieved accuracy $\geq 75.45\%$ for all languages. Moreover, except for English, they outperformed models that require translation or/and tokenization such as TransCNN (Word), a similar architecture that operates at word level and translated text (79.57% for English).

Multilanguage character embeddings are further developed in [55] but within an architecture (Word-Character CNN) that processes the text through two parallel CNN, one for words and the other for characters. The hypothesis is that words and character features provide complementary information. Outputs from both CNN are merged before being feed to a fully connected classifier. To achieve language-independence, the embedding layer is kept trainable. They used the same Twitter corpus as in [54]. The hybrid model yields a better performance ($\geq 77.13\%$) compared to pure word/character CNN such as [8] ($\geq 74.64\%$), [37] ($\geq 75.41\%$) and their former model UniCNNs [54] ($\geq 75.45\%$) for languages already studied in [54]. Interestingly, the two romance languages considered, Spanish (69.82%) and Portuguese (72.87%), had the worst performance.

Medrouk & Pappa [56] studied a similar architecture. It comprises a stack of CNN working as a feature extractor, i.e., an encoder, followed by polling and a fully connected predictor, but in this case, working at the n-gram level. To this point, CNN seems a popular choice within the domain in contrast to LSTM. The model is feed with reviews written in French, English, and Greek without any language indication. Empirical evaluation over a mix of contents from three languages yielded an F1 score of 88%. These results reinforce the assumption that the n-gram CNN can produce language-independent features capturing the local relations between words useful for multilingual polarity analysis.

Whether CNN and LSTM variants of the embedding-feature extractor-classifier architecture need extra pre-training hassle or additional complexity to handle

multilingual data is investigated by [57]. Experiments were conducted training monolingual and multilingual models, achieving accuracy over 90% for both types of networks working at the n-gram level. Moreover, the fact that the multilingual models behave as well as the monolingual ones, seems to confirm the hypothesis about their ability to extract rich features without distinction if processing single or multilingual datasets.

4.1.2. Aspect-based studies

Regardless of the promising results for SA at sentence level that achieves simpler architectures such as [57], it is not a surprise that for aspect level authors proposed more complex models.

The architecture (GRCNN-HBLSTM) proposed by [58] combines two word-level feature extractors. A BiLSTM encoding sentences that take as inputs the embeddings for the topics, the aspects, and the words. The original word embedding and a feature vector from a character CNN are combined through a gate mechanism to achieve language independence. The second encoder is a regional CNN that aims to preserve the temporal relationship between different sentences, also capturing some of the long-distance dependencies of the aspects. Both feature sets are feed to a sentence-level BiLSTM together with an attention mechanism. A softmax classifier handles the output of the last layer. In experiments using a subset of the dataset in [59] their full model yields an F1 score above 78.04% in all cases outperforming baselines such as the Hierarchical LSTM [31] ($\geq$ 78.04%). What is more important, they compared a version (CNN-HBLSTM) without the gate mechanism that achieves worsts results ($\geq$ 74.66%) and the highest variance among languages.

So far, we have reviewed the purely multilingual approaches for SA. We have contrasted very different approaches. However, at the sentence level, the common strategy is to learn features from a multilanguage set using CNN and feed a classifier module with those features. Unsurprisingly, for the aspect SA setup, authors embrace more complex architectures leveraging attention mechanisms and aspect embeddings. The next section is devoted to the cross-lingual

category.

4.2. Cross-lingual approaches

We categorized as cross-lingual the proposals where the focus is to leverage resource-rich language assets to extrapolate to a low-resource target, for example, through transfer learning. This is the core of the proposal in [60] for the cross-lingual projection of sentiment embeddings. Their custom architecture (Dual-Channel CNN) has one channel which works with word embeddings to extract features through convolutions. The other channel is similar but uses word sentiment embeddings which can boost the classification. Features computed from each channel are merged before being feed to a fully connected layer. It is worth noting that for low-resource languages, the sentiment embeddings can be projected from English. They evaluated their approach for English as the source and Spanish, Dutch, German, Russian, Italian, Czech, Japanese, and French. The induced embeddings lead to better results in 7 out of the 10 trials with accuracy over 79.3%.

While not common in cross-language SA, in [61] authors explored the architecture comprised of a feature extractor (BiLSTM) feed by embeddings followed by the classifier (dense layer) for transfer learning. First, they use a large dataset to train the whole model. Afterward, they fine-tuned in a small, labeled dataset from the target language, but only the embedding layer remains trainable. They trained using TripAdvisor reviews in English for the first stage and tweets in Greek for the second, being the results very sensible to the size of this dataset (accuracy drops from 91.7% to 73.2% as the dataset shrinks from 400 to 330). As the authors note, it will be interesting to study how a different degree of syntactic similarity between languages influences results.

Next works reviewed within the cross-lingual category explored the idea of using adversarial training to learn a set of language-independent features.

In [62] the authors delve into the synergies of microblog data in different languages from the same user to extract personalized language-specific or independent features to alleviate the lack of data in some sources. The architecture

has four components. First, an attention mechanism encoding users as feature vectors to propagate their individuality across the system. The second component are encoders $[\theta^1, \theta^2]$, one for each language, computing specific-language features. The third element is the language-independent encoder θ^G that is feed with sentences from both languages as well the user attention vector. Encoders are CNN over different n-gram representations of the sentences and an attention mechanism for the user-specific information. The classifier is softmax layer for each language with inputs from θ^G and θ^1 or θ^2. The last module is a Generative Adversarial Network (GA) that drives θ^G to a set of features useful for SA when combined with θ^1 or θ^2 but uncorrelated with the language of the input sentence. Experiments with Twitter and Sina Weibo compared monolingual baseline models and the proposal, trained with both languages at once. Results show an increase of the F1 score up to 2.12% respect the best monolingual ($\geq$ 79.85%).

Other work that leverages adversarial training to learn a set of language-independent but highly discriminatory features is [63]. The architecture (ADAN) uses the Deep Averaging Network (DAN) in [64] as a feature extractor with a Bilingual Word Embedding (BWE) [65] as an input layer. These features are feed to the classifier and to a language discriminator acting as an adversarial driving DAN to language-independent features. Empirical evaluation shows ADAN (accuracy $\geq$ 42.49%) improves in at least 6% over a version without the adversarial module (only DAN) trained with English to predict Chinese and Arab. It suggests that the adversarial mechanism is crucial for the results.

The adversarial mechanism is also critical in [66]. They build a cross-lingual word embedding using an Adversarial Auto Encoder (AAE) [67] feed from the outputs of two LSTM, one for each of the source and the target languages. On the top of this module sits a classifier based in Bidirectional Gated Recurrent Unit (BiGRU). Evaluating using Amazon comments with English as the source and Chinese and German as the target, the model (TL-AAE-BIGRU) achieved an F1 score $\geq$ 78.13%, about 3.44% better as average than the model without AEE ($\geq$ 73.25%). This result is consistent with [62, 63] who noticed the benefits

of the adversarial module too.

Instead of adversarial training, in [68] they opted to create universal embeddings by the combination of embeddings from high and low resources languages. The Universal SA (UniSent), involves the pre-training of two BiLSTM on English (labeled tweets), the alignment of low-resource language embeddings to the English ones with an unsupervised and domain-adversarial approach (MUSE [69]), and the fine-tuning on the low-resource languages validation sets applying an *Universal Embedding Layer*. This layer represents a word in a low-resource language by the weighted average of the most similar words to it in the English word embedding. The embeddings are feed to the classifier, a many-to-one LSTM layer. The experiments were carried in texts from OpeNER for Spanish and MultiBooked for Catalan. UniSent achieved F1 score $\geq$ 81.4% for the binary classification and $\geq$ 54.2%, outperforming even a version tested using translated texts ($\geq$ 74.0% and $\geq$ 40.6%).

The simpler architecture comprising LASER (Language-Agnostic SEntence Representations) toolkit[2] as a language-independent embedding, a feature extractor (CNN or BiLSTM) and the classifier is proposed by [70]. The multilingual embedding aims to drive the model to language-agnostic representations, being able to perform SA in languages different from the ones seen during training. Experiments training with Polish (F1 score 79.91%) to predict other languages seems to evidence this premise since in all cases, F1 was $\geq$ 77.96%. Also, that the setup LASER+BiLSTM is better in general.

A similar architecture is proposed by [71] but using Bilingual Bag-of-Words without Word Alignments (BilBOWA) and VecMap[3] as cross-lingual embedding. Experiments over English and Persian electronic product reviews evaluated different alternatives for the embeddings and the feature extractors (CNN, LSTM, CNN-LSTM, and LSTM-CNN). The VecMap+LSTM-CNN with dynamic embeddings, i.e., fine-tuning the embeddings with the training data,

[2] `https://github.com/facebookresearch/LASER`
[3] `https://github.com/artetxem/vecmap`

achieved the best results with an F1 score of 91.82%.

So far, we have reviewed cross-lingual approaches for SMA. Though there are very different perspectives to solve the problem, they also shared some commons ideas, as the projection of resources such as sentiment embeddings. The next section is committed to reviewing the works that addressed the code-switching setup.

4.3. Code-switching approaches

In this section, we examine the reports that tackled code-switching sentiment analysis. This setup poses challenges such as spelling variations, transliteration, informal grammar forms, and the scarcity of annotated data.

The underlying idea of [72] is to learn a sentiment feature space preserving the similarity of sentences in terms of the sentiment they convey. This enables us to measure the relatedness between code-switched content and labeled data from a resource-rich corpus. The Sentiment Analysis of Code-Mixed Text (SACMT) architecture uses a siamese BiLSTM with tri-gram embeddings as input and a fully connected layer on the top. They compared a model trained only with pairs of code-mixed text (Hindi-English) with an F1 score of 67.2%, to other trained with pairs of sentences one from English and other code-mixed (75.9%). These results suggest that in effect, the model benefits from the additional data provided by the English corpus.

Most of the research in code-mixed text has focused on the English-Hindi setup. One exception is the seminal work on code-mixed Bambara-French Facebook comments in [73]. They examined different variations of the base architecture with embeddings (at word or character level) as input, followed by a feature extractor (one of LSTM, BiLSTM, CNN, CNN-LSTM) and finally, the classifier. To mitigate the lack of pre-trained embeddings in Bambara, the model learns multilanguage embeddings from characters or words in the code-mixed corpus. The best performing model was a one-layer CNN model with an accuracy of 83.23%. The comparison between LSTM and CNN as feature extractors, where the latter-one yields better results, is coherent with a noticeable preference for

this type of model within the domain.

One problem when working with code-mixed data is the noise and the small size of datasets. To alleviate this, in [74] authors proposes to use n-gram embeddings instead of the subword ones suggested by [75]. Another novelty idea within the domain is the model that works as an ensemble of a Multinomial Naïve Bayes (MNB) and a recurrent neural network (LSTM or BiLSTM) classifier. They trained the MNB using both word-based 1 and 2-gram features while the neural networks models with the character 3-grams. The results, together with those reported by [75], suggest that for the LSTM network, the 3-grams representation is a better option (F1 score of 58.6%) over characters (51.1%). The subword level encoding achieved 65.2% but, the difference between the architectures can mislead the conclusions. Values for the ensemble (66.1%) show that this can be of benefit.

Authors of [76] assessed different versions of the architecture that combines sequentially one feature extractor and a classifier. The first one, a document to vector (Doc2Vec) layer whose output is feed to an SVM. The other three featured a FastText classifier, a BiLSTM, and a CNN with a softmax. The last two had a trainable word embedding layer before the NN. They curated a new corpus for Kannada-English (best results for the CNN model with an accuracy of 71.5%). Also, they evaluated using two available Hindi-English and, Bengali-English corpus [77] with the BiLSTM model achieving slightly better results, 60.22% and 72.2% respectively.

In [78] authors delve into whether to use characters, words, or subwords. Also, if projecting code-mixed text to a single feature space is a rich-enough representation for SA. Their architecture (CMSA) combines three parallel feature encoders before a classifier of four dense layers. The collective encoder aims to represent the overall sentiment of the sentences. It is based on a BiLSTM network whose end states are the features. The specific encoder is also a BiLSTM, but in this case, the intermediate states are also considered features through an attention mechanism. Both encoders take as input the output of a subword level CNN. The last one is a set of hand-crafted features to augment

the information supplied to the model. They evaluated the effect of removing some of the components. CMSA achieved an F1 score of 82.7%, better than the model only with the specific (80.1%), or only the collective (79.5%). It seems to support the hypothesis about the synergies of the different representations.

Another model combining different feature extractors is the proposed in [79]. Like [55], they have character-level and word-level feature encoders. The first one is inspired in [75], stacking a CNN followed by two LSTM layers. It aims to help with language independence, noise, and non-standard spelling. The other extractor stacks two LSTM layers. The concatenation of the two feature sets is the input for the classifier, a stack of two dense layers and a softmax. Their model achieved an F1 score of 66.13% over the Hindi-English corpus in [75] improving about 5.5% the baseline in [75]. As in [55], combining both types of feature extractors seems to lead to better results.

Recently in [80] authors evaluated several architectures (BiLSTM-CNN, Double BiLSTM, Attention-based), each one with and without GloVe, and BERT (Bidirectional Encoder Representations from Transformers) [81] - over tweets and Facebook comments for English-Hindi and English-Bengali. The same models were trained in a monolingual corpus to observe the effects of code-mixing. The best model for code-mixing, the Attention-based model with custom word embeddings, achieved an F1 score average of 0.66 and 0.67 for the monolingual setup. It is interesting that the performance of $\text{BERT}_{base-uncased}$, the best model for the monolingual with 0.77, decreased noticeably for the code-mixed (about 0.63).

The work [82] also relies on subword embeddings. The architecture (SAEKCS), similar to [75], includes a CNN layer on top of the embeddings to extract local dependencies. Its output is processed by a BiLSTM layer after max-pooling, to learn long-term relations. On the top, a fully connected layer acts as the classifier. They evaluated SAEKCS using Kannada-English code-switching YouTube comments with an accuracy of 77.6%. They also assessed a subword LSTM (64.8%) and a BiLSTM (55.9%), suggesting that the short-term dependencies encoded by the CNN greatly benefit the model.

Finally, we also have reviewed code-switching approaches for SMA. The more trending proposal is to use subword embeddings to allow guessing the meaning of unknown / out-of-vocabulary words. We have contrasted very different mixing languages, in the majority of cases, English mix. The next section is an overview of the deep learning implementations across the different setups.

4.4. An overview of the different deep learning implementations

Up to this point, we have reviewed a comprehensive corpus of research works that had leveraged deep learning for multilingual sentiment analysis. We had focused on the underlying hypothesis of the proposed approaches, highlighting what is common or different between the different solutions. Also, the results on the evaluation corpus.

In this section, we aim to make it easier for the reader to have a quick overview of the material we analyzed.

Table 1 lists the items of researches examined in this paper, summarizing the architectures they used in their best model (Proposed Model), baselines in their experiments, distinctive ideas (Proposed Approach), the classification categories, multilingualism level, languages, and information about the corpus.

The classification categories mainly will be divided into positive or negative (binary), also with a neutral class (ternary) or five classes[4].

[4]Sometimes authors remove the *neutral* class or add another, e.g., *ambivalent*.

Table 1: Multilingual Approaches Used in DSA.

Proposed Model*	**Baseline Model***	**Proposed Approach**	**C[1]**	**L[2]**	**Language**	**Corpus**	**Reference**
CNN	translate-single-lang-CNN, multi-lang-CNN, multi-lang-without-identification-CNN, Random Forest	Trained large amounts of data in various languages (is trained for every single language), in three phases: unsupervised, distant supervised, and supervised with multi-layer CNN	T	D & M	English, Italian, French, & German	Unlabeled tweets ($300M$), weakly-supervised data ($40 - 60M$), and annotated tweets ($71K$)	Deriu et al. (2017) [50]
CNN	**LSTM-Emb**, Conv-Emb, Conv-Emb-Freeze, Conv-Char & SVM	Cost-effective Character-based embedding and optimized convolutions	B	D	English, German, Portuguese & Spanish	Annotated tweets ($128K$, subset of $1.6M$)	Becker et al. (2017) [51]
CNN	LSTM-Emb, Conv-Emb, Conv-Emb-Freeze, Conv-Char & SVM	Character-level embeddings with few learnable parameters	B	D & M	English, German, Portuguese & Spanish	Annotated tweets ($128K$, subset of $1.6M$)	Wehrmann et al. (2017) [37]
CNN	word-Translation-CNN, char-Translation-CNN, 1-gram-SVM, 2-gram-SVM	Transformed characters into numbers corresponding UTF-8 decimal codes	B	D	English, Polish, German, Slovak, Slovenian & Swedish	Annotated tweets ($150K$, subset of $1.6M$)	Zhang et al. (2017) [54]
CNN	**SVM**	N-gram bilingual mixed (English-French) input text source (based on a Naïve approach)	B	D & M	French, English & Greek	Labeled restaurant balanced reviews (62.6 K)	Medrouk & Pappa (2017) [56]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
CNN	word-CNN, char-CNN, unicode-char-CNN, CuDNNLSTM (word and char)	Word-level & Character-level embeddings with two convolutional channels (one channel for each)	B	D	English, German, Portuguese, Spanish, Polish, Slovak, Slovenian & Swedish	Annotated tweets (193K, subset of 1.6M)	Zhang et al. (2017) [55]
BiLSTM	Average Skip-gram Vectors with LR & CNN-Subword-char-LSTM	Shared parameters of twin (siamese) networks with contrastive learning	T	M	English, Hindi	English annotated tweets (114K) & Hindi-English labeled sentences of Facebook posts (3.8K)	Choudhary et al. (2018) [72]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	**Baseline Model***	**Proposed Approach**	**C[1]**	**L[2]**	**Language**	**Corpus**	**Reference**
CNN	random initialized CNN, CNN + GloVe/FastText/Polyglot embeddings, regular CNN concatenate standard GloVe/FastText + multilingual sentiment embeddings (VADER, SocialSent or Amazon reviews), dual-channel-CNN + GloVe/FastText incorporates random initialized, Polyglot, VADER, SocialSent or static Amazon reviews embedding	Cross-lingual graph-based propagation (transfer-learning) from a rich source language with embeddings of supervised training on Amazon reviews to a dual-channel neural architecture	B	D	English, Spanish, Dutch, German, Russian, Italian, Czech, Japanese & French	Annotated movie reviews (12.2K, Rotten Tomatoes and AlloCine), labeled reviews ($20.8K$, TripAdvisor, and Amazon Fine Food) & labeled tweets ($4.8K$)	Dong & De Melo (2018) [60]
CNN	LSTM (one-layer and two-layer), BiLSTM (one-layer and two-layer), CNN-LSTM, NB & SVM	Dictionaries of character and word indexes to produce code-mixed character and word embedding for a single NN	T	M	Bambara & French (mixed)	Labeled Facebook comments ($17K$, subset of $74K$)	Konate & Du (2018) [73]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	**Baseline Model***	**Proposed Approach**	**C[1]**	**L[2]**	**Language**	**Corpus**	**Reference**
MNB + LSTM	Subword-LSTM	Ensemble of Multinomial Naïve Bayes with 1 and 2-gram features and many-to-one stacked LSTM over 3-gram encoding of sentences	T	M	English, Hindi	Hindi-English labeled sentences from Facebook posts (3.8K)	Jhanwar & Das (2018) [74]
LSTM	**CNN**	N-gram raw corpus-based input, without any preprocessing, translation, annotation nor additional knowledge features	B	D & M	French, English & Greek	Labeled restaurant and hotel balanced reviews (91.8K)	Medrouk & Pappa (2018) [57]
BiLSTM	Doc2Vec + SVM, FastText & CNN	Embedding with only distributed representation of the text	T	M	English, Bengali, Hindi & Kannada (English mixed)	Labeled sentences from Facebook comments (22.5*K*)	Shalini et al. (2018) [76]
BiLSTM	**SVM**	Learning new word embeddings based on limited training datasets and a pre-trained DNN exploiting transfer-learning from a rich source language with labeled data	B	D	English & Greek	Labeled TripAdvisor reviews (40*K*) & annotated tweets (480)	Stavridis et al. (2018) [61]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	**Baseline Model***	**Proposed Approach**	**C[1]**	**L[2]**	**Language**	**Corpus**	**Reference**
CNN + GAN + Attention-mechanism	SVM + Word2Vec, LSTM, CNN, NSC + UPA, UPNN	Combined CNN, GAN and user attention to learn specific and independent-language features from data with authorship information	B	D	English & Chinese	Annotated tweets ($48.1K$) & Weibo posts ($53.6K$)	Wang et al. (2018) [62]
GAN + DAN	DAN, mSDA, Machine Translation + DAN, CLD-KCNN, CLDFA-KCNN	Combined Bilingual Embedding (BWE), Deep Averaging Network (DAN) and adversarial training to learn independent-language features from a source language (English)	T, F	D	English, Chinese & Arab	English reviews from Yelp (700K), hotels reviews in Chinese ($20K$ labeled / $150K$ unlabeled) & tweets in Arab ($1.2K$ labeled)	Chen et al. (2018) [63]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
BiLSTM-CNN	CNN-hierarchical-BiLSTM, CNN-hierarchical-BiLSTM-gate-mechanism, LSTM, CNN, LSTM-attention-mechanism, CNN-attention-mechanism, RCNN-LSTM, hierarchical-LSTM, LSTM-sentences-relations	Word vector representation improvement based on gate mechanism, which obtains time-series relationship of different sentences in the comments through an RCNN, and gets the local features of the specific aspects in the sentence and the long-distance dependence in the whole comment through a hierarchical attention BiLSTM	B & T	D	English, Arabic, French & Chinese	Binary (400) and ternary ($3.7K$) labeled web reviews ($4.1K$)	Liu et al. (2019) [58]
BiLSTM-CNN	1-grams + 2-grams-SVM, 1-grams + 2-grams-NB-SVM, 1-grams + 2-grams-MNB, Tf-Idf-MNB, Lexicon Lookup, Char-LSTM, Subword-LSTM, FastText & SACMT	Hybrid architecture with subword level representations for the sentences, two parallel BiLSTM as Dual Encoder (Collective Encoder for overall sentiment and Specific Encoder with attention mechanism for subwords) and linguistic features network	T	M	English, Hindi	Hindi-English labeled sentences of Facebook posts ($3.8K$)	Lal et al. (2019) [78]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
BiLSTM	SVM-MONO, SVM-MT, ARTEXTE-SVM-based, ARTEXTE-Ensemble, BARISTA-SVM-based, BARISTA-Ensemble, BLSE, BLSE-Ensemble & BiLSTM-MT	Low-resource language embeddings + mapping function, joined with rich-resource language embedding through k-NN refinement, BiLSTM as encoder layer, then fully connected layer with softmax for prediction	F-1	D	English, Spanish & Catalan	English labeled tweets ($33.7K$), Spanish OpeNER ($1.3K$) & Catalan MultiBooked ($1K$)	Jabreel et al. (2019) [68]
Double LSTM	Double LSTM with several combinations of optimizers and loss functions & Subword-LSTM	Low resource + code-mixed corpus to train embeddings. Joint feature of sentences (subword + word levels), preceded by Double LSTM layer	T	M	English & Hindi (mixed)	Hindi-English labeled sentences of Facebook posts ($3.8K$)	Mukherjee (2019) [79]
LSTM-AAE-BiGRU	MT-SVM, MT-BiGRU & TL-BiGRU	Contextual word embeddings (Word2Vec+LSTM, source and target languages), AAE	B	D	English, Chinese & German	Amazon labeled documents ($28.9K$) and unlabeled documents ($80K$) for each pair of language category (books, DVD, music)	Shen et al. (2020) [66]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	Baseline Model*	Proposed Approach	C[1]	L[2]	Language	Corpus	Reference
BiLSTM + Attention-mechanism	NB, SMO (SVM), RF, BiLSTM-CNN, Double BiLSTM, GloVe + BiLSTM-CNN, GloVe + Double BiLSTM, GloVe + Attention-based-BiLSTM, BERT$_{base-uncased}$ & BERT$_{base-domain-uncased}$	Attention mechanism to extract such words that are important to the meaning of the sentence and aggregate the representation of those informative words to form the sentence vector; a sigmoid layer is used to predict the correct label	T	M	English, Hindi & Bengali (English mixed)	English, Bengali-English, Hindi-English annotated tweets (9.2K, 5.5K, 18.4K) & Hindi-English labeled sentences of Facebook posts (3.8K)	Jamatia et al. (2020) [80]
BiLSTM	LASER-CNN, FastText-BiLSTM & fastText-CNN	Transfer learning by LASER with (low-resource) language corpus, BiLSTM, then predict the sentiment of texts in other (high-resource) language	F + 1	D	Polish, Dutch, English, French, German, Italian, Portuguese, Russian & Spanish	Online medicine, hotels, school, products reviews (8.4K for each language)	Kanclerz et al. (2020) [70]
CNN-BiLSTM	NB, BiLSTM & Subword-LSTM	Three stages classification with subword embeddings + CNN-BILSTM: first positive or not, then negative or not, and then, computed classification matrix of them	T	M	English & Kannada (mixed)	Annotated YouTube comments (10.4K)	Chundi et al. (2020) [82]

continued . . .

Table 1: Multilingual Approaches Used in DSA (continued . . .).

Proposed Model*	**Baseline Model***	**Proposed Approach**	**C[1]**	**L[2]**	**Language**	**Corpus**	**Reference**
LSTM-CNN	BilBOWA + CNN, VecMap + CNN, BilBOWA + LSTM, VecMap + LSTM, BilBOWA + CNN-LSTM, VecMap + CNN-LSTM & BilBOWA + LSTM-CNN	Train bilingual embeddings (VecMap, on one high-resource and other low-resource language) and uses it on target language (low-resource), followed by a DL classifier for predict polarity	B	D	English & Persian	Binary (11K, Persian Digikala reviews), five categories (200K, English Amazon reviews)	Ghasemi et al. (2020) [71]

* **Bold**: model with best performance.

[1] Classification (**C**): **B** (Binary), **T** (Ternary), **F** (Five categories).

[2] Multilingualism Level (**L**): Document (**D**), Mix (**M**).

5. Discussion and future directions

In this section, we discuss the main findings of our study. We also highlight some unexplored topics that may hint at interesting directions for further research.

5.1. Languages and social media in MSA

The 24 studies we analyzed covered 23 different languages. In most cases English was the resource-rich language, except [70] and [73]. However, authors have explored synergies between different languages, as shown in Figure 1. Concerning the social media Twitter and Miscellaneous[5] account for most of the works. Figure 2 the data de-aggregated by language and social media.

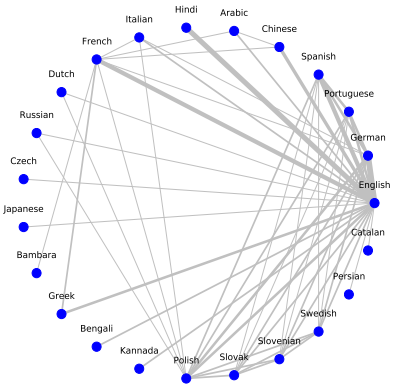


Figure 1: Synergies between languages. A link between two languages indicates that both has been used simultaneously in a model, as source or target language, or to learn multilingual feature spaces.

[5]Under this category, we considered works such as [58] which used the SemEval 2016 Task 5, and other corpora that aggregated text from different sources.

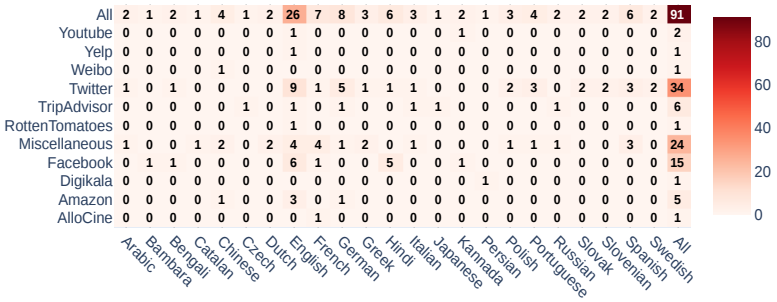


Figure 2: Languages vs Social Media

The next subsection is devoted to the analysis of the relations between the MSA setup and the architecture proposed to deal with the problem.

5.2. DL architectures for MSA

A comparison between the backbone of the different architectures suggests that in general, for multilanguage sentence-level SA, authors have explored a plainer architecture. It leverages trainable embeddings preceding the feature extractor and finally a classifier layer. Regardless, there is a wide range of alternatives in the design of the classifier, from a single BiLSTM [61] to parallel CNN [55]. Only one study focused on the aspect-level SA. So, it may be difficult to draw conclusions related to the best architectural decisions to tackle this problem. However, experiments in [58] suggest that the attention mechanism as well the aspect embedding plays a key role. Moreover, results from another work [83] show that the mere addition of attention to a simpler model such as LSTM or CNN does not lead to state-of-the-art results. This is indicative of the convenience of combining feature extractors at different levels for aspect-based SA.

For cross-lingual SA network designs are more diverse. However, two core ideas can be identified. The use of an adversarial module to drive the feature extractors to language-independent representations as in [62] or [63]. The other is leveraging trainable cross-lingual embeddings or even pre-trained ones such

as Facebook LASER [70].

Code-switching backbones tend to be simpler than cross-lingual, yet more elaborated than the multilingual ones. They can include parallel encoders at a character, word, or subword levels [78, 79] or implement ensemble models including deep neural networks and other classification techniques [74].

In general, a lot of effort has been devoted to compare feature extractors based on CNN, LSTM, or BiLSTM and different levels of embeddings. We analyzed the co-occurrences of the different types of networks, attention mechanisms, etc. within the same model, to visualize how authors have been using them. Edges in Figure 3 means that the two concepts at nodes have been used together in a model, the weight, how often this relationship has occurred. The graph shows how CNN is preferred along with hybrid architectures, where CNN + BiLSTM is the most studied with [73] and [76] reporting comparisons between models where they only changed this setting, resulting in better performance for the CNN models. Instead, results were equivalent in [57] and better for LSTM in [51]. Thus, seem to be no consensus about this subject.

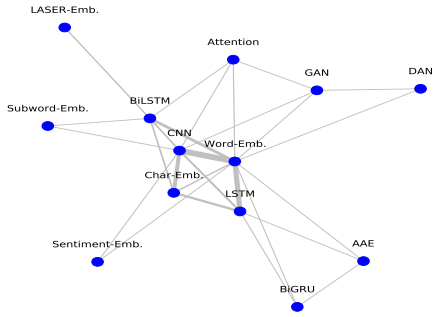


Figure 3: Neural Network architectures and its relations, out of 24, across reviewed papers.

5.2.1. Embedding approaches for MSA

Regardless of the domain, most authors rely on some embedding types, with a trend towards learning the embedding along the training process. There is

no common opinion about whether to work at a character, word, or subword. Authors such as [51] advocate character level embeddings due to their simplicity. However, they reported slightly better results for the word level case but in [37] shows that character level could help achieve a language-independence. In [73] they compared the character and word-based embeddings, with the latter yielding better results. Also, subword embeddings reported outperforming the character ones by [78].

Unsurprisingly we can find pre-trained subword multilingual embeddings, such as [84], which can be useful for MSA and SA for low-resource languages. However, there are another alternatives to be explored such as document level [76], a combination of different levels [55, 58] and even sentiment-driven embeddings [60], universal embeddings [68] or the use of some tools, such as LASER [70].

5.3. Future directions

Finally, we elaborate on the current state of research and provide a pathway for what can be done or needs to be done within the following few years.

Little-explored MSA levels:. Hitherto, Aspect-Based Sentiment Analysis has not widely been addressed using multilingual deep learning approaches. As [58] suggests, tackling this problem may require more complex architectures. Moreover, it needs to be studied if current proposals can handle mixing setups such as aspect-based code-switching.

MSA setup shift across time:. Figure 4 suggest a shift of the interest from multilingual to cross-lingual and code-switching approaches. In MSA this can be explained since initially most of the works focused on the multilingual setup evaluating many variations of the same design. Researchers could perceive this path as depleted. Also, the adoption of transformer-based architectures such as BERT [81] and even multilingual models such as *Multilingual BERT* [81] [6]

[6] `https://github.com/google-research/bert/blob/master/multilingual.md`

allows the researchers to focus on fine-tuning the models instead of training them from scratch with a multilingual corpus as has been typical for the multilingual setup.

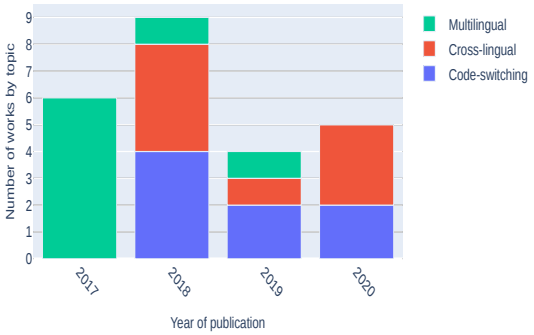


Figure 4: MSA approaches, out of 24 papers, across fours years.

Multilingual representations:. Multilingual embeddings and adversarial training seem to be the most common approaches to achieve multilingualism within the analyzed corpus. But there is not a common standpoint about the level of embedding to use or how a single model can encode a multilingual or language-agnostic feature space use-full for the downstream tasks. However, this debate seems to be shifted to the transformer-based architectures where different tokenizers are being considered [85]. Moreover, despite the success in training transformers in a multilingual corpus, recent studies suggest that there is a lot of room for improvement [86]. In this sense probably we will see an increased number of works studying the impact of the differences between languages and language families.

SA-specific representations:. For MSA, the aforementioned architectures would handle the specifics of this domain such as the code-switching or the aspect-based setup. It will be necessary to study if it is worth to couple techniques such as attention between different levels of representation, sentiment embeddings, or adversarial learning (e.g GAN-BERT [87]) into the aforementioned

architectures.

Low-resource languages and dialects:. Despite languages from different families has been studied (see Figure 2) the coverage is far from complete. Moreover, the steady interest in sentiment analysis, the lack so far of a universal approach, and the new opportunities [88, 89] would trigger the development of systems and corpora for SA in other languages. In this sense, we would see tailored solutions dealing with dialects and mixing languages. Besides India (native languages mixing with English) there are other large groups such as (a) Mexico and USA (Spanglish), (b) Brazil and its border countries, Portugal and Spain (Portuñol) (c) Paraguay (Jopara[7], Portuñol), to mention few cases. Nevertheless, the scarce of available corpora is a challenge for tackle these code-mixing tasks. For instance, in [80] we could observe that with an English mixing MSA setup, BERT had been unable to outperform the traditional DL models. Thus, substantial progress needs still to be made.

6. Conclusions

In this work, we reviewed 24 that studied 23 and 11 different languages and sources. The observed trend evidences the steady interest in this domain, so we expect to see this direction continue.

As regards the different MSA setups, the multilingual approach seems to be of decreased interest. However, aspect-based sentiment analysis is still an understudied domain and an open research field with a lot of scope for future works.

We highlighted the main ideas authors proposed to tackle the challenge that represents the lack of annotated data or to achieve language independent models. Despite state-of-the-art results in some cases, the simpler backbone comprising embeddings, a feature extractor, and a classifier seems to be unappropriated for more complex scenarios. Also, there are unsolved questions such as which type

[7]Mixing Guarani (an indigenous language) with Spanish.

of embedding captures better the particulars of MSA. We hint about future research directions, for example, if ideas such as contextualized embeddings, which have proven very useful in other tasks, can further improve MSA. Finally, although studies have covered very different languages such as Arabic, Chinese, or Hindi, the world is extraordinarily rich in languages, cultures, and ways of expressing feelings. Thus, better approaches need to be assessed or developed for new scenarios.

Acknowledgements

This research work has been partially funded by the Generalitat Valenciana (Conselleria dEducació, Investigació, Cultura i Esport) and the Spanish Government through the projects SIIA (PROMETEO/2018/089, PROMETEU/2018/089) and LIVING-LANG (RTI2018-094653-B-C22). We are also immensely grateful to David Vilares (Universidade da Coruña, Spain), for his recommendations provided a valuable orientation.

References

[1] S. L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artificial Intelligence Review 48 (4) (2017) 499–527.

[2] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Foundations and Trends® in Information Retrieval 2 (1–2) (2008) 1–135.

[3] S. Wang, C. D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proc. of the 50th annual meeting of the ACL: Short papers-volume 2, ACL, 2012, pp. 90–94.

[4] B. Liu, Sentiment analysis and opinion mining, Synthesis lectures on human language technologies 5 (1) (2012) 1–167.

[5] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.

[6] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proc. of the 2013 Conf. on empirical methods in NLP, 2013, pp. 1631–1642.

[7] C. Dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts, in: Proc. of COLING 2014, the 25th Int. Conf. on Computational Linguistics: Technical Papers, 2014, pp. 69–78.

[8] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. of the 2014 Conf. on Empirical Methods in NLP (EMNLP), ACL, Doha, Qatar, 2014, pp. 1746–1751.

[9] D. Tang, B. Qin, T. Liu, Deep learning for sentiment analysis: successful approaches and future challenges, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5 (6) (2015) 292–303.

[10] L. M. Rojas-Barahona, Deep learning for sentiment analysis, Language and Linguistics Compass 10 (12) (2016) 701–719.

[11] P. Singhal, P. Bhattacharyya, Sentiment analysis and deep learning: a survey, Center for Indian Language Technology, Indian Institute of Technology, Bombay.

[12] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, A. Rehman, Sentiment analysis using deep learning techniques: a review, Int J Adv Comput Sci Appl 8 (6) (2017) 424.

[13] D. Vilares, Compositional language processing for multilingual sentiment analysis, Ph.D. thesis, Universidade da Coruña (2017).

[14] L. J. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8.

[15] O. Habimana, Y. Li, R. Li, X. Gu, G. Yu, Sentiment analysis using deep learning approaches: an overview, Science China Information Sciences 63 (1) (2019) 1–36.

[16] R. Wadawadagi, V. Pagi, Sentiment analysis with deep neural networks: comparative study and performance assessment, ARTIFICIAL INTELLIGENCE REVIEW.

[17] H. Nankani, H. Dutta, H. Shrivastava, P. R. Krishna, D. Mahata, R. R. Shah, Multilingual sentiment analysis, in: Deep Learning-Based Approaches for Sentiment Analysis, Springer, 2020, pp. 193–236.

[18] J. M. Wiebe, R. F. Bruce, T. P. OHara, Development and use of a gold-standard data set for subjectivity classifications, in: Proc. of the 37th annual meeting of the ACL, 1999, pp. 246–253.

[19] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proc. of the ACL-02 Conf. on Empirical methods in NLP-Volume 10, ACL, 2002, pp. 79–86.

[20] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proc. of the 40th annual meeting on ACL, ACL, 2002, pp. 417–424.

[21] R. Feldman, Techniques and applications for sentiment analysis., Commun. ACM 56 (4) (2013) 82–89.

[22] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intelligent systems 28 (2) (2013) 15–21.

[23] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proc. of the 11th Int. Workshop on Semantic Evaluation, SemEval '17, ACL, Vancouver, Canada, 2017.

[24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Neural and Information Processing System (NIPS), 2013.

[25] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in NLP (EMNLP), 2014, pp. 1532–1543.

[26] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL-HLT, 2018, pp. 2227–2237.

[27] J. Yu, J. Jiang, Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification, ACL, 2016.

[28] H. Xu, B. Liu, L. Shu, P. S. Yu, Double embeddings and CNN-based sequence labeling for aspect extraction, in: Proc. of the 56th Annual Meeting of the ACL (Volume 2: Short Papers), ACL, 2018, pp. 592–598.

[29] B. Huang, Y. Ou, K. M. Carley, Aspect level sentiment classification with attention-over-attention neural networks, in: Int. Conf. on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Springer, 2018, pp. 197–206.

[30] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm, in: Thirty-second AAAI Conf. on artificial intelligence, 2018.

[31] S. Ruder, P. Ghaffari, J. G. Breslin, A hierarchical model of reviews for aspect-based sentiment analysis, in: Proc. of the 2016 Conf. on Empirical Methods in NLP, 2016, pp. 999–1005.

[32] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proc. of the 2015 Conf. on empirical methods in NLP, 2015, pp. 1422–1432.

[33] G. Rao, W. Huang, Z. Feng, Q. Cong, Lstm with sentence representations for document-level sentiment classification, Neurocomputing 308 (2018) 49 – 57.

[34] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proc. of the 32nd Int. Conf. on Int. Conf. on Machine Learning - Volume 37, ICML15, JMLR.org, 2015, p. 11801189.

[35] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, End-to-end adversarial memory network for cross-domain sentiment classification., in: IJCAI, 2017, pp. 2237–2243.

[36] B. Agarwal, R. Nayak, N. Mittal, S. Patnaik, Deep learning-based approaches for sentiment analysis (2020).

[37] J. Wehrmann, W. Becker, H. E. L. Cagnini, R. C. Barros, A character-based convolutional neural network for language-agnostic twitter sentiment analysis, 2017 Int. Joint Conf. on Neural Networks (IJCNN) (2017) 2384–2391.

[38] J. T. Zhou, S. J. Pan, I. W. Tsang, Y. Yan, Hybrid heterogeneous transfer learning through deep learning, in: Twenty-eighth AAAI Conf. on artificial intelligence, 2014.

[39] G. Zhou, Z. Zeng, J. X. Huang, T. He, Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks, in: Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, ACM, 2016, pp. 245–254.

[40] P. Singhal, P. Bhattacharyya, Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification, in: Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, 2016, pp. 3053–3062.

[41] Z. Wang, Y. Zhang, S. Lee, S. Li, G. Zhou, A bilingual attention network for code-switched emotion prediction, in: Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, 2016, pp. 1624–1634.

[42] S. Ghosh, S. Ghosh, D. Das, Sentiment identification in code-mixed social media text, ArXiv abs/1707.01184.

[43] G. I. Ahmad, J. Singla, N. Nikita, Review on sentiment analysis of indian languages with a special focus on code mixed indian languages, in: 2019

Int. Conf. on Automation, Computational and Technology Management (ICACTM), IEEE, 2019, pp. 352–356.

[44] E. Tromp, Multilingual sentiment analysis on social media, Lap Lambert Academic Publ, 2012.

[45] C. Banea, R. Mihalcea, J. Wiebe, Multilingual sentiment and subjectivity analysis, Multilingual NLP 6 (2011) 1–19.

[46] I. S. V. Roncal, Multilingual sentiment analysis in social media, Ph.D. thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea (2019).

[47] F. Steiner-Correa, M. I. Viedma-del Jesus, A. Lopez-Herrera, A survey of multilingual human-tagged short message datasets for sentiment analysis tasks, Soft Computing 22 (24) (2018) 8227–8242.

[48] M. A. Abdullah, Deep learning for sentiment and emotion detection in multilingual contexts, Ph.D. thesis, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; ltima actualizacin - 2019-10-18 (2018).

[49] B. Ay Karakuş, M. Talo, İ. R. Hallaç, G. Aydin, Evaluating deep learning models for sentiment classification, Concurrency and Computation: Practice and Experience 30 (21) (2018) e4783.

[50] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, M. Jaggi, Leveraging large amounts of weakly supervised data for multi-language sentiment classification, in: Proc. of the 26th Int. Conf. on world wide web, Int. World Wide Web Conf. Steering Committee, 2017, pp. 1045–1052.

[51] W. Becker, J. Wehrmann, H. E. Cagnini, R. C. Barros, An efficient deep neural architecture for multilingual sentiment analysis in twitter, in: The Thirtieth Int. Flairs Conf., 2017.

[52] I. Mozetič, M. Grčar, J. Smailović, Multilingual twitter sentiment classification: The role of human annotators, PloS one 11 (5) (2016) e0155036.

[53] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[54] S. Zhang, X. Zhang, J. Chan, Language-independent twitter classification using character-based convolutional networks, in: Int. Conf. on Advanced Data Mining and Applications, Springer, 2017, pp. 413–425.

[55] S. Zhang, X. Zhang, J. Chan, A word-character convolutional neural network for language-agnostic twitter sentiment analysis, in: Proc. of the 22nd Australasian Document Computing Symposium, ACM, 2017, p. 12.

[56] L. Medrouk, A. Pappa, Deep learning model for sentiment analysis in multilingual corpus, in: Int. Conf. on Neural Information Processing, Springer, 2017, pp. 205–212.

[57] L. Medrouk, A. Pappa, Do deep networks really need complex modules for multilingual sentiment polarity detection and domain classification?, in: 2018 Int. Joint Conf. on Neural Networks (IJCNN), IEEE, 2018, pp. 1–6.

[58] G. Liu, X. Huang, X. Liu, A. Yang, A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network, The Computer Journal.

[59] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al., Semeval-2016 task 5: Aspect based sentiment analysis, in: Proc. of the 10th Int. workshop on semantic evaluation (SemEval-2016), 2016, pp. 19–30.

[60] X. Dong, G. De Melo, Cross-lingual propagation for deep sentiment analysis, in: Thirty-Second AAAI Conf. on Artificial Intelligence, 2018.

[61] K. Stavridis, G. Koloniari, E. Keramopoulos, Deriving word embeddings using multilingual transfer learning for opinion mining, in: 2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conf. (SEEDA_CECNSM), IEEE, 2018, pp. 1–6.

[62] W. Wang, S. Feng, W. Gao, D. Wang, Y. Zhang, Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning, in: Proc. of the 2018 Conf. on Empirical Methods in NLP, 2018, pp. 338–348.

[63] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, K. Weinberger, Adversarial deep averaging networks for cross-lingual sentiment classification, in: Proc. of the 2018 Conf. on Empirical Methods in NLP, 2018, pp. 557–570.

[64] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, Deep unordered composition rivals syntactic methods for text classification, in: Proc. of the 53rd Annual Meeting of the ACL and the 7th Int. Joint Conf. on NLP (Volume 1: Long Papers), 2015, pp. 1681–1691.

[65] W. Y. Zou, R. Socher, D. Cer, C. D. Manning, Bilingual word embeddings for phrase-based machine translation, in: Proc. of the 2013 Conf. on Empirical Methods in NLP, 2013, pp. 1393–1398.

[66] J. Shen, X. Liao, S. Lei, Cross-lingual sentiment analysis via aae and bigru, in: 2020 Asia-Pacific Conf. on Image Processing, Electronics and Computers (IPEC), IEEE, 2020, pp. 237–241.

[67] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial autoencoders`arXiv:1511.05644`.

[68] M. Jabreel, N. Maaroof, A. Valls, A. Moreno, Unisent: Universal sentiment analysis system for low-resource languages., in: CCIA, 2019, pp. 387–396.

[69] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, H. Jgou, Word translation without parallel data, in: Int. Conf. on Learning Representations, 2018.

[70] K. Kanclerz, P. Miłkowski, J. Kocoń, Cross-lingual deep neural transfer learning in sentiment analysis, Procedia Computer Science 176 (2020) 128–137.

[71] R. Ghasemi, S. A. Ashrafi Asli, S. Momtazi, Deep persian sentiment analysis: Cross-lingual training for low-resource languages, Journal of Information Science (2020) 1–14.

[72] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of code-mixed languages leveraging resource rich languages, in: 19th Int. Conf. on Computational Linguistics and Intelligent Text Processing (CICLing-2018), 2018.

[73] A. Konate, R. Du, Sentiment analysis of code-mixed bambara-french social media text using deep learning techniques, Wuhan University Journal of Natural Sciences 23 (3) (2018) 237–243.

[74] M. Gopal Jhanwar, A. Das, An ensemble model for sentiment analysis of hindi-english code-mixed data, arXiv preprint arXiv:1806.04450.

[75] A. Joshi, A. Prabhu, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text, in: Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, 2016, pp. 2482–2491.

[76] K. Shalini, H. B. Ganesh, M. A. Kumar, K. Soman, Sentiment analysis for code-mixed indian social media text with distributed representation, in: 2018 Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2018, pp. 1126–1131.

[77] B. G. Patra, D. Das, A. Das, Sentiment analysis of code-mixed indian languages: an overview of sail_code-mixed shared task@ icon-2017, arXiv preprint arXiv:1803.06745.

[78] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing sentiment from code-mixed text, in: Proc. of the 57th Annual Meeting of the ACL: Student Research Workshop, 2019, pp. 371–377.

[79] S. Mukherjee, Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features, in: 2019 IEEE 16th India Council Int. Conf. (INDICON), IEEE, 2019, pp. 1–4.

[80] A. Jamatia, S. Swamy, B. Gambäck, A. Das, S. Debbarma, Deep learning based sentiment analysis in a code-mixed english-hindi and english-bengali social media corpus, Int. Journal on artificial intelligence tools 29 (5).

[81] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805v2.

[82] R. Chundi, V. R. Hulipalled, J. Simha, Saekcs: Sentiment analysis for english–kannada code switchtext using deep learning techniques, in: 2020 Int. Conf. on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), IEEE, 2020, pp. 327–331.

[83] Y. Zhu, X. Gao, W. Zhang, S. Liu, Y. Zhang, A bi-directional lstm-cnn model with attention for aspect-level text classification, Future Internet 10 (12) (2018) 116.

[84] B. Heinzerling, M. Strube, BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages, in: N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[85] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art nlp, in: Proc. of the 2020 Conf. on Empirical Methods in NLP: System Demonstrations, ACL, Online, 2020, pp. 38–45.

[86] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: Proc. of the 57th Annual Meeting of the ACL, ACL, Florence, Italy, 2019, pp. 4996–5001.

[87] D. Croce, G. Castellucci, R. Basili, GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples, in: Proc. of the 58th Annual Meeting of the ACL, ACL, Online, 2020, pp. 2114–2119.

[88] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowledge and Information Systems 60 (2) (2019) 617–663.

[89] A. Zunic, P. Corcoran, I. Spasic, Sentiment analysis in health and well-being: Systematic review, JMIR Med Inform 8 (1) (2020) e16023.

A.5 On the logistical difficulties and findings of Jopara Sentiment Analysis

- Agüero-Torales, M., Vilares, D., & López-Herrera, A. (2021, June). On the logistical difficulties and findings of Jopara Sentiment Analysis. In Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching (pp. 95-102). *Association for Computational Linguistics (North American Chapter)*. doi: `http://dx.doi.org/10.18653/v1/2021.calcs-1.12`.
  - Status: Published.
  - Google Scholar metric (h5 2021): index 105, median 195.
  - Subject Category: Engineering & Computer Science, Computational Linguistics. Ranking 3/20.

A.5.1 Summary

According to what was studied in the previous section (§A.4), we were put into practice what we reviewed, applying several machine learning approaches (from traditional to more advanced, i.e., transformer-based) to a specific low-resource language: Guarani, as well as to its mixture with Spanish (i.e., Jopara).

This paper summarized the effort of collecting Guarani-Spanish (i.e., Jopara) text data from Twitter for sentiment analysis and reported the results provided by different classification systems. The created corpus, so-called JOSA, is a valuable resource of low-resourced Guarani-dominant texts with certain limitations, like unbalanced classes due to the nature of the corpus. Note that collected tweets pertain to a few Twitter accounts that usually tweet in Guarani, due to the scarce of tweets in Guarani-dominant. A brief discussion was conducted on the classification results provided by the different models. In addition, an error analysis was performed, which provided further information on the performance of the classifiers.

As mentioned above, this paper described an approach to sentiment analysis on Jopara, a code-switching language between Guarani and Spanish. Three types of models was tested to perform sentiment analysis on this corpus of Jopara tweets: (i) traditional machine learning models (Naive Bayes and SVM), (ii) Bidirectional LSTM and CNN neural models (fed both with non-contextualized word and character embeddings) and (iii) transformer-based models (Spanish-BERT 'BETO' [285], Multilingual BERT 'mBERT'[8] [12] and Cross-lingual Language Model - XLM [38]). The best results were obtained with BETO model, despite being pre-trained in Spanish (not in Guarani or Jopara) and the dataset, Guarani-dominant.

We presented the first Guarani-dominant dataset for sentiment analysis with a well-structured description of the challenges that we had to face to train models and even create low-resources languages datasets.

[8] `https://github.com/google-research/bert/blob/master/multilingual.md`

On the logistical difficulties and findings of Jopara Sentiment Analysis

Marvin M. Agüero-Torales
DECSAI, University of Granada
Granada, Spain
maguero@correo.ugr.es

David Vilares
Universidade da Coruña, CITIC
A Coruña, Spain
david.vilares@udc.es

Antonio G. López-Herrera
DECSAI, University of Granada
Granada, Spain
lopez-herrera@decsai.ugr.es

Abstract

This paper addresses the problem of sentiment analysis for Jopara, a code-switching language between Guarani and Spanish. We first collect a corpus of Guarani-dominant tweets and discuss on the difficulties of finding quality data for even relatively easy-to-annotate tasks, such as sentiment analysis. Then, we train a set of neural models, including pre-trained language models, and explore whether they perform better than traditional machine learning ones in this low-resource setup. Transformer architectures obtain the best results, despite not considering Guarani during pre-training, but traditional machine learning models perform close due to the low-resource nature of the problem.

1 Introduction

Indigenous languages have been often marginalized, an issue that is reflected when it comes to design natural language processing (NLP) applications, where they have been barely studied (Mager et al., 2018). One of the places where this is greatly noticed is Latin America, where the dominant languages (Spanish and Portuguese) coexist together with hundreds of indigenous languages such as Guarani, Quechua, Nahuatl or Aymara.

In this context, the Guarani language plays a particular role. It is an official language in Paraguay and Bolivia. Besides, it is spoken in other regions, e.g. Corrientes (Argentina) or Mato Grosso do Sul (Brazil), alongside with their official languages. Overall, it has about 8M speakers. Its coexistence with other languages, mostly Spanish, has contributed to its use in code-switching setups (Muysken, 1995; Gafaranga and Torras, 2002; Matras, 2020) and led to Jopara, a code-switching between Guarani and Spanish, with flavours of Portuguese and English (Estigarribia, 2015).

Despite its official status, there is still few NLP resources developed for Guarani and Jopara. Adelali et al. (2006) developed a parallel Spanish-English-Guarani corpus for machine translation. Similarly, Chiruzzo et al. (2020) developed a Guarani-Spanish parallel corpus aligned at sentence-level. There are also a few online dictionaries and translators from Guarani to Spanish and other languages.[1] Beyond machine translation, Maldonado et al. (2016) released a corpus for Guarani speech recognition that was collected from the web; and Rudnick (2018) presented a system for cross-lingual word sense disambiguation from Spanish to Guarani and Quechua languages. There also are a few resources for PoS-tagging and morphological analysis of Guarani, such as the work by Hämäläinen (2019) and Apertium;[2] and also for parsing, more specifically for the Mbyá Guarani variety (Dooley, 2006; Thomas, 2019), under the Universal Dependencies framework.

In the context of sentiment analysis (SA; Pang et al., 2002; Liu, 2012), and more particularly classifying the polarity of a text as positive, negative or neutral, we are not aware of any previous work; with the exception of (Ríos et al., 2014). They presented a sentiment corpus for the Paraguayan Spanish dialect, which also includes words in English and Portuguese. However, there were few, albeit relevant, words of Guarani (70) and Jopara[3] (10), in comparison to the amount of the ones in Spanish (3,802) (Ríos et al., 2014, p. 40, Table II). Overall, SA has focused on rich-resource languages for which data is easy to find, even when it comes to code-switching setups (Vilares

[1] https://gn.wiktionary.org/, https://es.duolingo.com/dictionary/Guarani/, https://www.paraguay.gov.py/traductor-guarani, https://www.iguarani.com/, https://glosbe.com/gn, and Mainumby (Gasser, 2018).

[2] https://github.com/apertium/apertium-grn

[3] Tokens that mix n-grams of characters from Guarani and Spanish, e.g.: '*I understand*' would be '*entiendo*' (es), '*ahechakuaa*' (gn) and '*aentende*' (jopara).

et al., 2016), maybe with a few exceptions such as English code-switched with languages found in India (Sitaram et al., 2015; Patra et al., 2018; Chakravarthi et al., 2020). In this context, although some previous work has developed multilingual lexicons and methods (Chen and Skiena, 2014; Vilares et al., 2017); for languages such as Guarani and other low-resource cases (where web text is scarce), it is hard to develop NLP corpora and systems.

Contribution Our contribution is twofold. First, we collect a corpus for polarity classification of Jopara tweets, which mixes Guarani and Spanish languages, being the former the dominating language in the corpus. We also discuss on the difficulties that we had to face when creating such resource, such as finding enough Twitter data that shows sentiment and contains a significant amount of Guarani terms. Second, we train a set of neural encoders and also traditional machine learning models, in order to have a better understand of how old versus new models perform in this low-resource setup, where the amount of data matters.

2 JOSA: The Jopara Sentiment Analysis dataset

In what follows, we describe our attempts to collect Jopara tweets. Note that ideally we are interested in tweets that are as Guarani as possible. However, Guarani is intertwined with Spanish, and thus we have focused on Jopara, aiming for Guarani-dominant tweets, in contrast to Ríos et al. (2014). We found interesting to report failed attempts to collect such data, since the proposed methods would most likely work to collect data in rich resource languages. We hope this can be helpful for other researchers interested in developing datasets for low-resource languages in web environments.

In this line, Twitter does not allow to automatically crawl Guarani tweets, since it is not included in its language identification tool. To overcome this, we considered two alternatives: (i) using a set of Guarani keywords (§2.1), and (ii) scrapping Twitter accounts that mostly tweet in Guarani (§2.2).

2.1 Downloading tweets using Guarani keywords - An unsuccessful attempt.

As the Twitter real-time streamer can deal with a limited number of keywords, we consider 50 different keywords which are renewed every 3 hours, and used them to sample tweets. To select such keywords, we considered two options:

1. *Dictionary-based keywords*: We used 5.1K Guarani terms from a Spanish-Guarani word-level translator.[4] We then downloaded 2.1M tweets and performed language identification with three tools: (i) `polyglot`,[5] (ii) `fastText` (Joulin et al., 2016) and (iii) `textcat`.[6] We assume that the text was Guarani if at least one of them classified the text as Guarani. After this, we got 5.3K tweets. Next, a human annotator was in charge of classifying such subset, obtaining that only 150 tweets, over the initial set of 2.1M samples, were prone to be Guarani-dominant.

2. *Corpus-based keywords*: We first merged two Guarani datasets[7] (Scannell, 2007), that were generated from web sources and included biblical passages, wiki entries, blog posts or tweets, among other sources. From there, we selected 550 terms, including word uni-grams and bi-grams with 100 occurrences or more. Again, we downloaded tweets using the keywords and collected 7M of tweets, but after repeating the language identification phase of step 1, we obtained a marginal amount of tweets that were Guarani-dominant.

Limitations This approach suffered from a low recall when it came to collect Guarani-dominant tweets, while similar approaches have worked when collecting data for rich-resource languages, where a few keywords were enough to succesfully download tweets in the target language (Zampieri et al., 2020). In this context, even if tweets contained a few Guarani terms, there were other issues: (i) words that have the same form in Spanish and Guarani such as '*mano*' ('*hand*' and '*to die*'), (ii) loanwords,[8] such as '*pororo*' ('*popcorn*') or '*chipa*' (traditional Paraguayan food, non-translatable); (iii) or simply tweets where the majority of the content was written in Spanish. Overall, this has been a problem experienced in other low-resource setups (Hong et al., 2011; Kreutz and Daelemans,

[4] https://github.com/SENATICS/traductor-espanhol-guarani

[5] https://polyglot.readthedocs.io/en/latest/Detection.html

[6] https://www.nltk.org/_modules/nltk/classify/textcat.html

[7] BCP-47 *gn* and *gug* codes.

[8] Frequent in Paraguay and border countries (Pinta, 2013).

2020), so we decided instead to look for alternatives to find Guarani-dominant tweets.

2.2 Downloading tweets from Guarani accounts - A successful attempt.

In this case, we crawled Twitter accounts that usually tweet in Guarani.[9] We scrapped them, and obtained more than 23K Guarani and Jopara tweets from a few popular users (see Appendix A.1). Using the same Guarani language identification approach as in 1, we obtained $8,716$ tweets. To eliminate very similar tweets that could contaminate the dataset, we removed tweets with a similarity greater than 60%, according to the Levenshtein distance. After applying this second cleaning step, we obtained a total of 3,948 tweets.

The dataset was then annotated by two native speakers of Guarani and Spanish. They were asked to: (i) determine whether the tweet was strictly written in Guarani, Jopara or other language (i.e., if the tweet did not have any words in Guarani); and determine whether the tweet was positive, neutral or negative. For sentiment annotations consolidation, we proceeded similarly to the SemEval-2017 Task 4 guidelines (Rosenthal et al., 2017, § 3.3).[10] We then filtered the corpus by language, including only those labeled as Guarani or Jopara, to ensure the samples are Guarani-dominant. This resulted into $3,491$ tweets.

Limitations Although this second approach is successful when it comes to collect a reasonable amount of Guarani-dominant tweets, it also suffers from a few limitations. For instance, the first part of Table 1 shows that due to the nature of the crawled Twitter accounts (who tweet about events, news, announcements, greetings, ephemeris, tweets to encourage the use of Guarani, etc.), there is a tendency to neutral tweets. Also, as the number of selected accounts was small, the number of discussed topics might be limited too. We comment on this a bit further in the Appendix A.1.

Balanced and unbalanced versions As we are interested in identifying sentiment in Jopara tweets, we also created a balanced version of JOSA. Note that unbalanced settings are also interesting and might reflect real-world setups. Thus, we will report results both on the unbalanced and balanced setups. More particularly, we split each corpus into training (50%), development (10%), and test (40%). We show the statistics in Table 1.

For completeness, in Table 2 we show for the balanced corpus the top five most frequent terms (we only consider content tokens) for Guarani, Spanish and some language-independent tokens, such emoticons. This was done based on a manual annotation of a Guarani-Spanish native speaker.

Version	Total	Positive	Neutral	Negative
Unbalanced	$3,491$	349	$2,728$	414
Balanced	$1,526$		763	

Version		Train	Development	Test
Unbalanced	$3,491$	$1,745$	349	$1,397$
Balanced	$1,526$	763	152	611

Table 1: JOSA statistics and splits for the unbalanced/balanced versions.

Category	#Terms	Most frequent
Guarani	$4,336$	guaranime, ñe'ẽ, mba'e, guarani, avei
Spanish	$1,738$	paraguay, guaraní, no, es, día
Other*	$1,440$	alcaraz, su, rt, juan, francisco
Mixing	368	guaraníme, departamento-pe, castellano-pe, castellanope, twitter-pe
Emojis	112	🇵🇾 😂 👏 xD :)

*We include reserved words, proper nouns, acronyms, etc.

Table 2: Frequent terms for the balanced JOSA.

3 Models

Due to the low-resource setup, we run neural models and pre-trained language models, but also other machine learning models, such as complement naïve Bayes (CNB) and Support Vector Machines (SVMs) (Hearst et al., 1998), since they are less data hungry, and could help shed some light about the real effectiveness of neural models on Jopara texts. In all cases, the selection of the hyperparameters was done over a small grid search based on the dev set. We report the details in the Appendix A.2.

Naïve Bayes and SVMs We tokenized the tweets[11] and represented them as a 1-hot vector of unigrams with a TF-IDF weighting scheme. We used Pedregosa et al. (2011) for training.

Neural networks for text classification We took into account neural networks that process in-

[9] We followed http://indigenoustweets.com/gn/. We did not use an external human annotator as in 1, since the crawled accounts tend to tweet in Guarani.

[10] We obtained a slight agreement following Cohen's kappa metric (Artstein and Poesio, 2008).

[11] We used the TweetTokenizer from the NLTK library.

put tweets as a sequence of token vector representations. More particularly, we consider both long short-term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and convolutional neural networks (CNN) (LeCun et al., 1995), as implemented in NCRF++ (Yang and Zhang, 2018). Although the former are usually more common in many NLP tasks, the latter have also showed traditionally a good performance on sentiment analysis (Kalchbrenner et al., 2014).

For the input word embeddings, we tested: (i) randomly initialized word vectors, following an uniform distribution, (ii) and pre-trained non-contextualized representations and more particularly, FastText's word vectors (Bojanowski et al., 2017) and BPEmb's subword vectors (including the multilingual version, which supports Guarani) (Heinzerling and Strube, 2018). In both cases, we also concatenate a second word embedding, computed through a char-LSTM (or CNN).

Pre-trained language models We also fine-tuned recent contextualized language models on the JOSA training set. We tested BERT (Devlin et al., 2019) including: (i) beto-base-uncased (a Spanish BERT) (Cañete et al., 2020), and (ii) multilingual bert-base-uncased (mBERT-base-uncased, pre-trained on 102 languages). We also tried more recent variants of multilingual BERT, in particular XLM (Lample and Conneau, 2019). Note that BERT models use a wordpiece tokenizer (Wu et al., 2016) to generate a vocabulary of the most common subword pieces, rather than the full tokens, and that in the case of the multilingual models, none of the language models used considered Guarani during pre-training.

4 Experiments

Reproducibility The baselines and tweet IDs[12] are available at https://github.com/mmaguero/josa-corpus.

We run experiments for the unbalanced and balanced versions of JOSA, evaluating the macro-accuracy (to mitigate the impact of the neutral class in the unbalanced setup). Table 3 shows the comparison. Note that all models, even the non-deep-learning models, only use raw word inputs and do not consider any additional information or hand-crafted features,[13] yet they obtained results that are in line with those of more recent approaches.

	Corpus	
Model	**Unbalanced**	**Balanced**
CNB	0.50	0.55
SVM	0.55	0.54
^{C}CNN-^{W}BiLSTM	0.45	0.57
^{C}BiLSTM-^{W}CNN	0.49	0.53
$_{BPEmb,gn}$ ^{C}CNN-^{W}BiLSTM	0.46	0.53
$_{BPEmb,gn}$ ^{C}BiLSTM-^{W}CNN	0.42	0.50
$_{BPEmb,es}$ ^{C}CNN-^{W}BiLSTM	0.45	0.52
$_{BPEmb,es}$ ^{C}BiLSTM-^{W}CNN	0.45	0.50
$_{BPEmb,m}$ ^{C}CNN-^{W}BiLSTM	0.47	0.52
$_{BPEmb,m}$ ^{C}BiLSTM-^{W}CNN	0.43	0.48
$_{FastText,gn}$ ^{C}CNN-^{W}BiLSTM	0.46	0.53
$_{FastText,gn}$ ^{C}BiLSTM-^{W}CNN	0.42	0.51
$_{FastText,es}$ ^{C}CNN-^{W}BiLSTM	0.46	0.52
$_{FastText,es}$ ^{C}BiLSTM-^{W}CNN	0.46	0.46
BETO$_{base,uncased}$	**0.64**	**0.64**
mBERT$_{base,uncased}$	0.55	0.58
XLM-MLM-TLM-XNLI-15	0.46	0.49

^{C}Encodes character sequence. ^{W}Encodes word sequence.
Pre-trained embeddings are represented with a prefix together with their language ISO 639-1 code (except for m: multilingual).

Table 3: Experimental results on JOSA, both on the balanced and unbalanced setups.

With respect to the experiments with CNNs and BiLSTMs encoders, we tested different combinations using character representations, which output is first concatenated to a second external word vector (as explained in §3), and then fed to the encoder. Among those, the model that used a character-level CNN and a word-level BiLSTM encoder obtained the best results. Still, the difference with respect to traditional machine learning models is small. We hypothesize this might be due to the low-resource nature of the task. Finally, the pre-trained language models that use transformers architectures, in particular BETO, obtain overall the best results, despite not being pre-trained on Guarani. We believe this is partly due to the presence of Spanish words in the corpora and also to the cross-lingual abilities that BERT model might explode, independently of the amount of word overlap (Wang et al., 2019).

Error analysis on the balanced version of JOSA Figure 1 shows the confusion matrices for a representative model of each machine learning family (based on the accuracy): (i) CNB, (ii) the best BiLSTM-based model (CNN-BiLSTM), and (iii) Spanish BERT (BETO). There seems to be different tendencies in the miss-classifications that different models make. For instance, CNB tends to over-classify tweets as negative, while both deep

[12] Contact the authors for more details.

[13] In order to keep an homogeneous evaluation setup.

learning models show a more controlled behaviour when predicting this class. Although for the three models neutral tweets seem to be the easiest to identify, both deep learning models are clearly better at it. Finally, when it comes to identify positive tweets, BETO seems to show the overall best performance. These different tendencies indicate that an ensemble method could be beneficial for low-resource setups such as the ones that JOSA represent, since the models seem to be complementary to certain extent. In this context, we would like to explore this line of work in the future, following previous studies such as Jhanwar and Das (2018), which showed the benefits of combining different machine learning models for Hindi-English code-switching SA.

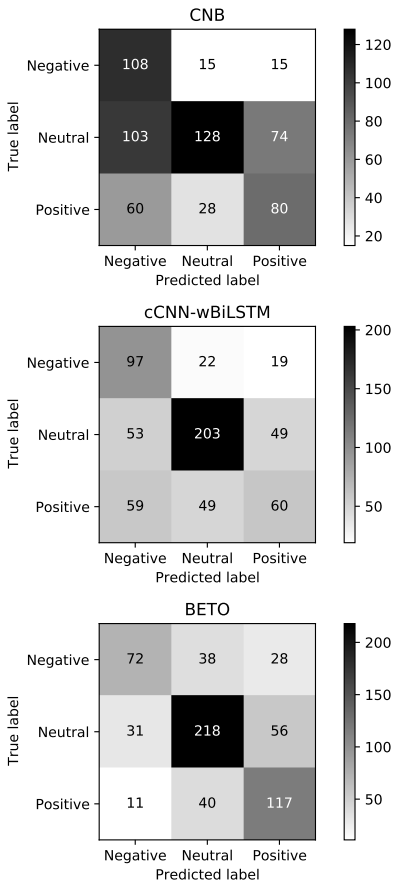


Figure 1: Confusion matrix for the balanced version of JOSA and the predictions of a representative member of each machine learning family: CNB, a BiLSTM-based model and Spanish BERT (BETO).

5 Conclusion

This paper explored sentiment analysis on Jopara, a code-switching language that mixes Guarani and Spanish. We collected the first Guarani-dominant dataset for sentiment analysis, and described some of the challenges that we had to face to create a collection where there is a significant number of Guarani terms. We then built several machine learning (naïve Bayes, SVMs) and deep learning models (BiLSTMs, CNNs and BERT-based models) to shed light about how they perform on this particular low-resource setup. Overall, transformers models obtain the best results, even if they did not consider Guarani during pre-training. This poses interesting questions for future work such as how cross-lingual BERT abilities (Wang et al., 2019) can be exploited for this kind of setups, but also how to improve language-specific techniques that can help process low-resource languages efficiently.

Acknowledgements

We thank the annotators that labelled JOSA. We also thank ExplosionAI for giving us access to the Prodigy annotation tool[14] with the Research License. DV is supported by a 2020 Leonardo Grant for Researchers and Cultural Creators from the FBBVA.[15] DV also receives funding from MINECO (ANSWER-ASAP, TIN2017-85160-C2-1-R), from Xunta de Galicia (ED431C 2020/11), from Centro de Investigación de Galicia 'CITIC', funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program) by grant ED431G 2019/01.

References

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski. 2006. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,"Visions for the Future of Machine Translation*, pages 1–9.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[14] https://prodi.gy/

[15] The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert A Dooley. 2006. Léxico guarani, dialeto mbyá, com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística. e referências. *Cuiabá: Summer Institute of Linguistics*.

Bruno Estigarribia. 2015. Guarani-spanish jopara mixing in a paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222.

Joseph Gafaranga and Maria-Carme Torras. 2002. Interactional otherness: Towards a redefinition of codeswitching. *International Journal of Bilingualism*, 6(1):1–22.

Michael Gasser. 2018. Mainumby: un ayudante para la traducción castellano-guaraní. *arXiv preprint arXiv:1810.08603*.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lichan Hong, Gregorio Convertino, and Ed Chi. 2011. Language matters in twitter: A large scale study. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Madan Gopal Jhanwar and Arpita Das. 2018. An ensemble model for sentiment analysis of hindi-english code-mixed data. *CoRR*, abs/1806.04450.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665.

Tim Kreutz and Walter Daelemans. 2020. Streaming language-specific Twitter data with optimal keywords. In *Proceedings of the 12th Web as Corpus Workshop*, pages 57–64, Marseille, France. European Language Resources Association.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Diego Manuel Maldonado, Rodrigo Villalba Barrientos, and Diego P Pinto-Roa. 2016. Eñe'ẽ: Sistema de reconocimiento automático del habla en guaraní. In *Simposio Argentino de Inteligencia Artificial (ASAI 2016)-JAIIO 45 (Tres de Febrero, 2016)*.

Yaron Matras. 2020. *Language contact*. Cambridge University Press.

Pieter Muysken. 1995. Code-switching and grammatical theory. *The bilingualism reader*, pages 280–297.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: an overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Justin Pinta. 2013. Lexical strata in loanword phonology: Spanish loans in guaraní. Master's thesis, The University of North Carolina at Chapel Hill.

Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623.

Adolfo A Ríos, Pedro J Amarilla, and Gustavo A Giménez Lugo. 2014. Sentiment categorization on a creole language with lexicon-based and machine learning techniques. In *2014 Brazilian Conference on Intelligent Systems*, pages 37–43. IEEE.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Alexander James Rudnick. 2018. *Cross-Lingual Word Sense Disambiguation for Low-Resource Hybrid Machine Translation*. Ph.D. thesis, Indiana University.

Kevin P Scannell. 2007. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Dinkar Sitaram, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. 2015. Sentiment analysis of mixed language employing hindi-english code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 271–276. IEEE.

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France. Association for Computational Linguistics.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2016. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A Alonso. 2017. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Appendix

A.1 Twitter user accounts

We scraped the following Twitter user accounts and mentions: @ndishpy, @chereraugo, @Pontifex_grn, @lenguaguarani, @enga_paraguayo, @SPL_Paraguay, @rubencarlosoje1, as well as some keywords: '*guaranime*', '*avañe'ẽme*', '*remiandu*', '*#marandu*', '*reikuaavéta*', '*hesegua*', '*reheguápe*', '*rejuhúta*'.

Note that accounts such as @Pontifex_grn, @SPL_Paraguay and @lenguaguarani belong to influential people and organizations. For instance, the first belongs to Pope Franscico, the second to the Secretariat of Linguistic Policy of Paraguay, and the third is the account of the General Director of the 'Athenaeum of the Guarani Language and Culture'. On the other hand, the terms '*marandu*' (news) and '*remiandu*' (feeling, sense) are related to news, where the first term means 'news' or 'to report' and the second is the name of a Paraguayan newspaper section[16] that publishes in Guarani.

A.2 Hyperparameters search and implementation details

To set the machine learning baselines, two standard classifiers were chosen: a variant of Naïve Bayes, Complement Naïve Bayes (CNB) (Rennie et al., 2003) to correct the 'severe assumptions' made by the standard Multinomial NB classifier; and Support Vector Machine (SVM) using weighted classes, to mitigate the effect of unbalanced classes. For the CNB, we set $\alpha = 0.1$ and considered only unigrams, except for the balanced version, where the combined use of unigrams and bigrams showed more robust results. To train the SVMs, we tested different values for the kernels: the `sigmoid` kernel obtained the best results for the unbalanced version of JOSA, and the `poly` kernel obtained the best results for the balanced version.

We used the NCRF++ Neural Sequence Labeling Toolkit (Yang and Zhang, 2018) to train our deep learning models and the Hugging face package (Wolf et al., 2020) for the transformer-based models. Table 4 shows the hyper-parameters used to train these models, both for the unbalanced and balanced corpus. The pre-trained embeddings used for Spanish, Guarani (and also the multilingual ones) have 300 dimensions. Finally, we trained the CNN and BiLSTM models for 20 epochs with a batch size of 10, and the transformer-based models were trained for up to 40 epochs relying on early stopping (set to 3). To train the models we used a NVIDIA Tesla T4 GPU with 16GB.

Parameter	Options
Sklearn	
TF-IDF-Lowercase	[True, False]
TF-IDF-n-grams	[(1,1) - (3,3)]
SVM-Kernel	[poly, sigmoid, linear, rbf]
CNB-alpha	[1.0, 0.1]
NCRF++	
Optimizer	[Adam, AdaGrad, SGD]
Avg. batch loss	[True, False]
Learning rate	[5e-5 - 0.2]
Char hidden dim.	[100, 200, 400, 800]
Word hidden dim.	[50, 100, 200]
Momentum	[0.0, 0.9, 0.95, 0.99]
LSTM Layers	[1, 2]
Hugging Face	
Eval. steps	[200]
Eval. strategy	[steps]
Disable tqdm	[False]
Eval. batch size	[16, 32]
Train batch size	[16, 32]
Learning rate	[2e-5 - 3e-5*]
Dropout	[0.1 - 0.6]
Epoch	[30 - 40]
Weight decay	[0.0 - 0.3]

*Except for the multilingual models, where 5e-5 was necessary to converge.

Table 4: Hyperparameters for the training of the models, both for the unbalanced and balanced corpus.

[16] https://www.abc.com.py/especiales/remiandu/

Appendix B

Evaluation and Annotation Guides

The appendix describes the quantitative analysis for topic modeling used in chapter 4, as well as the annotation guidelines of the Guarani-dominant Jopara corpora used in chapter 5 and chapter 6, and described in chapter 3.

B.1 Annotation guides for Guarani and Jopara corpora

In this part, we present the annotation guidelines of the Guarani-dominant Jopara corpora for sentiment analysis. Note that the guidelines are written in Spanish, but also have an English translation.

B.1.1 Sentiment analysis

- Agüero-Torales, Marvin M. (2021). Guarani and Jopara sentiment analysis - Mini-annotation guides. ResearchGate. doi:http://dx.doi.org/10.13140/RG.2.2.21174.63048.

Análisis de sentimientos en guaraní y jopará

Mini-guías de anotación

Guarani and Jopara sentiment analysis

Mini-annotation guides

Marvin M. Agüero-Torales.[1]

[1] maguero@correo.ugr.es, Department of Computer Science and Artificial Intelligence, University of Granada, February 2021.

Tareas

(For English translation please see below).

El propósito de esta guía es facilitar la determinación del sentimiento en un texto expresado en lenguaje natural (guaraní, jopará o guaraní-español) y contenidas especialmente en el contexto de una red social. A su vez, la identificación del idioma en el que está escrito el texto.

Para la anotación utilizaremos la herramienta prodigy[2], en la que tendremos las categorías para cada tarea a la hora de anotar.

Idioma

Determinar, en qué idioma está escrito el texto.

- (language) **guarani**: es para identificar si está en guaraní al menos el 90% del texto;
- (language) **guarani-other/s_language/s**: una mezcla del guaraní con uno o varios idiomas (español: Jopará, inglés, portugués, etc.);
- (language) **other**: texto no contiene guaraní (e.g., español "puro", o algún otro idioma).

Ejemplos

- **guarani**: "kuña he'i voi"
- **guarani-other/s_language/s**: "Che tavyeteko hese pero che añembotavy gua'u chugui😈"
- **other**: "@oscar_marandu Penal para olimpia"

Sentimiento

Determinar, qué sentimiento **predomina** en el texto.

- [sentiment] **strongly_positive**: es muy positivo el tweet, es decir, tiene muchos sentimientos positivos o pocos pero muy marcados.
- [sentiment] **weakly_positive**: el tweet es medianamente positivo, es positivo pero no en gran medida.
- [sentiment] **neutral**: Ni positivo ni negativo, es objetivo, es decir, dice las cosas tal cual sin emitir sentimiento.
- [sentiment] **weakly_negative**: el tweet es medianamente negativo, es negativo pero no en gran medida.
- [sentiment] **strongly_negative**: es muy negativo el tweet, es decir, tiene muchos sentimientos negativos o pocos pero muy cargados.

[2] https://prodi.gy/demo

Ejemplos

- **strongly_positive**: “el maestro ... la leyenda Roger Federer rey del juego de revés uno de sus mejores tiros”
- **weakly_positive**: “El domingo con Leeds United es como la cena del sábado, muy agradable”
- **neutral**: “Apple lanza una nueva actualización de su sistema operativo”
- **weakly_negative**: “Los refugiados se enfrentan a dificultades”
- **strongly_negative**: “las actualizaciones de mi teléfono Apple... son una mierda”

Task

---------- English translation ----------

The purpose of this guide is to facilitate the determination of the sentiment in a text expressed in natural language (Guarani, Jopará, or Guarani-Spanish) and contained especially in the context of a social network. In turn, the identification of the language in which the text is written.

For the annotation, we will use the prodigy[3] tool, in which we will have the categories for each task at the time of annotation.

Language

Determine in which language the text is written.

- (language) **guarani**: is to identify if at least 90% of the text is in Guarani;
- (language) **guarani-other/s_language/s**: a mixture of Guarani with one or more languages (Spanish: Jopará, English, Portuguese, etc.);
- (language) **other**: the text does not contain Guarani (e.g., "pure" Spanish, or some other language).

Examples

- **guarani**: "kuña he'i voi"
- **guarani-other/s_language/s**: "Che tavyeteko hese pero che añembotavy gua'u chugui😈"
- **other**: "@oscar_marandu Penal para olimpia"

Sentiment

Determine which sentiment **predominates** in the text.

- [sentiment] **strongly_positive**: the tweet is very positive, i.e. it has a lot of positive sentiments or few but very strong ones.
- [sentiment] **weakly_positive**: the tweet is moderately positive, it is positive but not to a great extent.
- [sentiment] **neutral**: Neither positive nor negative, it is objective, i.e., it tells it like it is without giving off feelings.
- [sentiment] **weakly_negative**: the tweet is moderately negative, it is negative but not to a great extent.
- [sentiment] **strongly_negative**: the tweet is very negative, i.e., it has a lot of negative feelings or few but very charged.

[3] https://prodi.gy/demo

Examples

- **strongly_positive**: “el maestro ... la leyenda Roger Federer rey del juego de revés uno de sus mejores tiros”
- **weakly_positive**: “El domingo con Leeds United es como la cena del sábado, muy agradable”
- **neutral**: “Apple lanza una nueva actualización de su sistema operativo”
- **weakly_negative**: “Los refugiados se enfrentan a dificultades”
- **strongly_negative**: “las actualizaciones de mi teléfono Apple... son una mierda”

B.1.2 Multi-annotated sentiment analysis: Emotion detection, offensive language identification, and humor detection

- Agüero-Torales, Marvin M. (2021). Emotion detection, offensive language identification, and humor detection in Guarani and Jopara - Mini-annotation guides. ResearchGate. doi: http://dx.doi.org/10.13140/RG.2.2.12786.02244.

Detección de emociones, identificación de lenguaje ofensivo y/o soez y detección de humor en guaraní y jopará

Mini-guías de anotación

Emotion detection, offensive language identification, and humor detection in Guarani and Jopara

Mini-annotation guides

Marvin M. Agüero-Torales.[1]

[1] maguero@correo.ugr.es, Department of Computer Science and Artificial Intelligence, University of Granada, July 2021.

Tareas

(For English translation please see below).

El propósito de esta guía es facilitar la determinación de la emoción en un texto expresado en lenguaje natural (guaraní, jopará o guaraní-español) y contenidas especialmente en el contexto de una red social. A su vez, la identificación del idioma en el que está escrito el texto, como la presencia o no de un tono gracioso, divertido o jocoso, o bien, de un tono ofensivo o que emplea lenguaje vulgar o soez.

Para la anotación utilizaremos la herramienta prodigy[2], en la que tendremos las categorías para cada tarea a la hora de anotar.

Idioma

Determinar, en qué idioma está escrito el texto.

- (language) **guarani**: texto que está escrito en guaraní paraguayo, o bien, "puro";
- (language) **guarani-other/s_language/s**: una mezcla del guaraní con uno o varios idiomas (español: Jopará, inglés, portugués, etc.);
- (language) **other**: texto no contiene guaraní (español "puro", algún otro idioma).

Ejemplos

- **guarani**: "kuña he'i voi"
- **guarani-other/s_language/s**: "Che tavyeteko hese pero che añembotavy gua'u chugui😈"
- **other**: "@oscar_marandu Penal para olimpia"

Emoción

Determinar, qué emoción **predomina** en el texto.

- {emotion} 😀 **happy**: texto que demuestra el sentir felicidad, el estar contento, tener placer o satisfacción;
- {emotion} 😢 **sad**: texto que demuestra el sentir o mostrar dolor, infelicidad, tristeza, lamento;
- {emotion} 😠 **angry**: texto que demuestra el sentir o mostrar una fuerte molestia, disgusto u hostilidad, estar lleno de ira, enojo, enfado;
- {emotion} 😶 **other**: texto que no demuestra ninguna de las anteriores emociones o no presenta emoción (emoción neutra).

Ejemplos

- **happy**: "Che tavyeteko hese pero che añembotavy gua'u chugui😈"

[2] https://prodi.gy/demo

- **sad**: "Ambyasyvoi rohayhu hague ingrata paloma blanca"
- **angry**: "kuña he'i voi"
- **other**: "@elaguelodice Nde tujaaaama reho.."

Lenguaje ofensivo, racista, vulgar, obsceno y/o grosero

Determinar, si se usa lenguaje ofensivo o no.

- [offensive] 😡 **it's offensive?**: texto que contiene lenguaje inapropiado, insultos o amenazas;
- [offensive] 🤗 **non-offensive**: texto que no es ofensivo ni profano.

Ejemplos

- **it's offensive?**: "kuña he'i voi"
- **non-offensive**: "Che tavyeteko hese pero che añembotavy gua'u chugui🌚"

Lenguaje divertido, gracioso, chistoso y/o humorístico

Determinar, si hay humor (si es gracioso/divertido) o no.

- <funny> 😂 **it's funny?**: texto que contiene lenguaje divertido, gracioso, chistes o ñe'ẽnga;
- <funny> 😐 **non-funny**: texto que no es divertido ni gracioso.

Ejemplos

- **it's funny?**: "Che tavyeteko hese pero che añembotavy gua'u chugui🌚"
- **non-funny**: "kuña he'i voi"

Task

---------- English translation ----------

The purpose of this guide is to facilitate the determination of emotion in a text expressed in natural language (Guarani, Jopará, or Guarani-Spanish) and contained especially in the context of a social network. In turn, the identification of the language in which the text is written, such as the presence or absence of a humorous or funny tone, or an offensive tone or one that uses vulgar or coarse language.

For annotation, we will use the prodigy[3] tool, in which we will have the categories for each annotation task.

Language

Determine in which language the text is written.

- (language) **guarani**: text that is written in Paraguayan Guarani, or "pure";
- (language) **guarani-other/s_language/s**: a mixture of Guarani with one or more languages (Spanish: Jopará, English, Portuguese, etc.);
- (language) **other**: text does not contain Guarani ("pure" Spanish, some other language).

Examples

- **guarani**: "kuña he'i voi"
- **guarani-other/s_language/s**: "Che tavyeteko hese pero che añembotavy gua'u chugui😈"
- **other**: "@oscar_marandu Penal para olimpia"

Emotion

Determine which emotion **predominates** in the text.

- {emotion} 😀 **happy**: text that demonstrates feeling happiness, being happy, having pleasure or satisfaction;
- {emotion} 😢 **sad**: text that shows feeling or showing pain, unhappiness, sadness, sorrow;
- {emotion} 😠 **angry**: text that demonstrates feeling or showing strong discomfort, disgust or hostility, being full of anger, angry, annoyed;
- {emotion} 😶 **other**: text that shows none of the above emotions or no emotion (neutral emotion).

Examples

- **happy**: "Che tavyeteko hese pero che añembotavy gua'u chugui😈"
- **sad**: "Ambyasyvoi rohayhu hague ingrata paloma blanca"

[3] https://prodi.gy/demo

- **angry**: "kuña he'i voi"
- **other**: "@elaguelodice Nde tujaaaama reho.."

Offensive, racist, vulgar, obscene, and/or rude language

Determine whether or not offensive language is used.

- [offensive] 🤬 **it's offensive?**: text that contains inappropriate language, insults or threats, etc..;
- [offensive] 🤗 **non-offensive**: text that is neither offensive nor profane.

Examples

- **it's offensive?**: "kuña he'i voi"
- **non-offensive**: "Che tavyeteko hese pero che añembotavy gua'u chugui🌚"

Funny, humorous language and/or jokes

Determine, if there is humor (if it is funny) or not.

- <funny> 😂 **it's funny?**: text containing funny, humorous language, jokes or “ñe'ẽnga”;
- <funny> 😐 **non-funny**: text that is neither funny nor humorous.

Examples

- **it's funny?**: "Che tavyeteko hese pero che añembotavy gua'u chugui🌚"
- **non-funny**: "kuña he'i voi"

B.2 Quantitative analysis for topic modeling

- Agüero-Torales, Marvin M. (2021). Quantitative Analysis for Topic Modeling - Mini-evaluation framework and mini-guides. ResearchGate. doi:http://dx.doi.org/10.13140/RG.2.2.18639.20647.

Quantitative analysis for Topic Modeling

Mini-evaluation framework and mini-guides

Marvin M. Agüero-Torales.[1]

[1] maguero@correo.ugr.es, Department of Computer Science and Artificial Intelligence, University of Granada, February 2021.

Human evaluation (topic modeling)

Objective

To create an evaluation framework.

Please referer to section 4.4, p. 186 of the paper for more details:

Agüero-Torales, M. M., Vilares, D., & López-Herrera, A. G. (2021). Discovering topics in Twitter about the COVID-19 outbreak in Spain. Procesamiento del Lenguaje Natural, 66, 177-190. URL:http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6333.

Method

We take X (20) random topics from among all the periods, where each topic is described by its most representative words and sentences, to make a human evaluation.

Annotator must:

(i) determine if see a clear topic,

(iii) for each representative phrase (tweet), determinate if think it belongs to that topic and,

(iii) the same for each of the 8 most representative words.

Guidelines

File info

There are two Google-sheets files (**random_topics_phrase**[2] and **random_topics_word**[3]), they both have three common columns:

Raw1: This is the raw tweet (phrase) most representative of the given topic.

Raw2: This is the second raw tweet (phrase) most representative of the given topic.

Raw3: This is the third raw tweet (phrase) most representative of the given topic.

DS_Keywords: These are the 8 most representative words of the given topic, calculated from its discriminative score.

period_topicNumber_total: Topic ID, which is composed of the period covered by the given topic, the number that corresponds to it (ranging from 0 to 9), and the total number of tweets that make up this topic.

In addition, **random_topics_word**, has the "**word**" column, where the 8 most representative words of the given topic are listed, one by one, row by row.

Columns to annotate

In the first file, **random_topics_phrase**, the annotation columns are as follows:

Is it a clear topic ?: Select **Y**: Yes, if you see a clear topic based on its most representative words and the phrase (tweet). Otherwise **N**: No.

[2] https://drive.google.com/file/d/1-kjGaSo28uNsZBpPWdWRTXzkNNMpocTc/view?usp=sharing
[3] https://drive.google.com/file/d/1-mdoTGXMzmWZMxENg6jqp1x1Mh2xpvqs/view?usp=sharing

Do the tweets belong to this topic ?: Select **Y**: Yes, if you think the tweet belongs to that topic. Otherwise **N**: No.

In the second file, random_topics_word, the only column to annotate is the following:

Does the word belong to this topic ?: Select **Y**: Yes, if you think the word belongs to that topic. Otherwise **N**: No.

Appendix C

Hyperparameters search and implementation details

In this part, we present details on the implementation and hyperparameter search of the machine learning models used in chapters 5 and 6 of this thesis. As well as details about the scraped Twitter accounts in the subsection 3.2.2.

Twitter user accounts In subsection 3.2.2, we scraped the following Twitter features:

- User accounts and mentions: *@ndishpy*, *@chereraugo*, *@Pontifex_grn*, *@lenguaguarani*, *@enga_paraguayo*, *@SPL_Paraguay*, *@rubencarlosoje1.*
- Keywords: '*guaranime*', '*avañe'ẽme*', '*remiandu*', '*#marandu*', '*reikuaavéta*', '*hesegua*', '*reheguápe*', '*rejuhúta*'.

Note that accounts such as *@Pontifex_grn*, *@SPL_Paraguay* and *@lenguaguarani* belong to influential people and organizations. For instance, the first belongs to Pope Franscico, the second to the Secretariat of Linguistic Policy of Paraguay, and the third is the account of the General Director of the 'Athenaeum of the Guarani Language and Culture'. On the other hand, the terms '*marandu*' (news) and '*remiandu*' (feeling, sense) are related to news, where the first term means 'news' or 'to report' and the second is the name of a Paraguayan newspaper section[1] that publishes in Guarani.

Hyperparameters search in chapter 5 To set the machine learning baselines, two standard classifiers were chosen: a variant of Naïve Bayes, Complement Naïve Bayes (CNB) [328] to correct the 'severe assumptions' made by the standard Multinomial NB classifier; and Support Vector Machine (SVM) using weighted classes, to mitigate the effect of unbalanced classes. For the CNB, we set $\alpha = 0.1$ and considered only unigrams, except for the balanced version, where the combined use of unigrams and bigrams showed more robust results. To train the SVMs, we tested different values for the kernels: the `sigmoid` kernel obtained the best results for the unbalanced version of JOSA, and the `poly` kernel obtained the best results for the balanced version.

We used the *NCRF++ Neural Sequence Labeling Toolkit* [283] to train our deep learning models and the *Hugging face* package [77] for the transformer-based models. Table C.1 shows the hyper-parameters used to train these models, both for the unbalanced and balanced corpus. The pre-trained embeddings used for Spanish, Guarani (and also the multilingual ones) have 300 dimensions. Finally, we trained the CNN and BiLSTM models for 20 epochs with a batch size of 10, and the transformer-based models were trained for up to 40 epochs

[1] `https://www.abc.com.py/especiales/remiandu/`

Parameter	Options
Sklearn	
TF-IDF-Lowercase	[True, False]
TF-IDF-n-grams	[(1,1) - (3,3)]
SVM-Kernel	[poly, sigmoid, linear, rbf]
CNB-alpha	[1.0, 0.1]
NCRF++	
Optimizer	[Adam, AdaGrad, SGD]
Avg. batch loss	[True, False]
Learning rate	[5e-5 - 0.2]
Char hidden dim.	[100, 200, 400, 800]
Word hidden dim.	[50, 100, 200]
Momentum	[0.0, 0.9, 0.95, 0.99]
LSTM Layers	[1, 2]
Hugging Face	
Eval. steps	[200]
Eval. strategy	[steps]
Disable tqdm	[False]
Eval. batch size	[16, 32]
Train batch size	[16, 32]
Learning rate	[2e-5 - 3e-5*]
Dropout	[0.1 - 0.6]
Epoch	[30 - 40]
Weight decay	[0.0 - 0.3]

*Except for the multilingual models, where 5e-5 was necessary to converge.

Table C.1: Hyperparameters for the training of the models on JOSA, both for the unbalanced and balanced corpus.

relying on early stopping (set to 3). To train the models we used a NVIDIA Tesla T4 GPU with 16GB.

Parameter	Options
NCRF++ [283]	
Optimizer	[Adam, AdaGrad, SGD]
Avg. batch loss	[True, False]
Learning rate	[2e-5, 5e-4, 0.015, 0.16, 0.1, 0.2]
Char hidden dim.	[50, 100, 200, 400]
Word hidden dim.	[50, 100, 200, 400]
Momentum	[0.0, 0.9, 0.95, 0.99]
LSTM Layers	[1, 2]
Dropout	[0.0, 0.5]
Iteration	[20,30,40,50]
Hugging Face [77]	
Eval. steps	[200]
Eval. strategy	[steps]
Disable tqdm	[False]
Eval. batch size[a]	[8, 16, 32]
Train batch size[a]	[8, 16, 32]
Learning rate	[2e-5 - 5e-5][b]
Dropout	[0.1 - 0.6]
Epoch[a]	[10 - 50]
Weight decay	[0.0 - 0.3]

[a]Except for the JOSA dataset, where Epoch = [30 - 40] and batch size = [16, 32] were used for comparison purposes.

[b]Except for the original BETO model, where 3e-5 was enough to converge.

Table C.2: Hyperparameters for the training of the models, both for the JOTAD and JOSA dataset.

Hyperparameters search in chapter 6 Table C.2 shows the hyper-parameters used to train all proposed models, using (i) the *NCRF++ Neural Sequence Labeling Toolkit* [283] to train our deep learning models and (ii) the *Hugging Face* package [77] for the transformer-based models.

In the case of the models with transformers, the selection of the hyper-parameters was done with a Bayesian hyper-parameter search method using `W&B` platform[2] [329] over the dev set. This method chooses parameters to optimize the probability of improvement according to the relationship between such parameters and the model metric, which in our case was the macro-accuracy. Instead, we trained the CNN and BiLSTM models with a batch size of 10 and the selection of the hyper-parameters was done with a random hyper-parameter search method. Finally, we trained the models for up to 50 epochs. The transformer-based models relying on early stopping criteria (set to 3). To train all the models we used a GPU with 12GB.

[2] `https://docs.wandb.ai/guides/sweeps`