Rocío Jiménez-Briones
(ed.)

Approaches to Knowledge Representation and Language

Granada, 2024

Colección indexada en la MLA International Bibliography desde 2005

EDITORIAL COMARES

INTERLINGUA
000

ENVÍO DE PROPUESTAS DE PUBLICACIÓN:

Las propuestas de publicación han de ser remitidas (en archivo adjunto, con formato PDF) a alguna de las siguientes direcciones electrónicas: anabelen.martinez@uco.es, psgines@ugr.es

Antes de aceptar una obra para su publicación en la colección INTERLINGUA, ésta habrá de ser sometida a una revisión anónima por pares. Para llevarla a cabo se contará, inicialmente, con los miembros del comité científico asesor. En casos justificados, se acudirá a otros especialistas de reconocido prestigio en la materia objeto de consideración.

Los autores conocerán el resultado de la evaluación previa en un plazo no superior a 60 días. Una vez aceptada la obra para su publicación en INTERLINGUA (o integradas las modificaciones que se hiciesen constar en el resultado de la evaluación), habrán de dirigirse a la Editorial Comares para iniciar el proceso de edición.

Maquetación: …

Polígono Juncaril • C/ Baza, parcela 208 • 18220 Albolote (Granada) • Tlf.: 958 465 382
https://www.comares.com • E-mail: libreriacomares@comares.com
https://www.facebook.com/Comares • https://twitter.com/comareseditor
https://www.instagram.com/editorialcomares

ISBN: … • Depósito legal: Gr. …/2024

Impresión y encuadernación: COMARES

Sumario

Acknowledgements IX

Introduction 1

Rocío Jiménez-Briones
Avelino Corral Esteban

I
KNOWLEDGE REPRESENTATION FOR SOCIAL SENSING AND LANGUAGE

Chapter 1.—Exploring problems through social media: The case of Beach Quality 11

Carlos Periñán-Pascual

Chapter 2.—Description of social problems by means of schemas related to the Income domain in the DIAPASON platform 27

Ángel Felices-Lago

Chapter 3.—Integrating corpus methodology into the construction of an intelligent crowdsensing system 45

Pedro Ureña Gómez-Moreno

Chapter 4.—Social media detection of texts on the social problem of violence against women within the multimodal intelligent system ALLEGRO 63

Ángela Alameda Hernández

Chapter 5.—Exploring the creation of synthetic corpora of negative communicative functions for the task of communicative function identification 77

Nicolás José Fernández-Martínez

Chapter 6.—Negation as a polarity shifter in English: An NLP approach 95

Yolanda Blázquez-lópez

Chapter 7.—A hybrid conceptual model using statistical and linguistic methods for pre-screening memory through language and cognition 111

Kulvinder Panesar
María Beatriz Pérez Cabello de Alba

II
KNOWLEDGE REPRESENTATION FOR COGNITION AND LANGUAGE

Chapter 8.—The Spanish implicational constructicon: A case study from the perspective of Cognitive Linguistics 135

María Sandra Peña-Cervel
Carla Ovejas-Ramírez

Chapter 9.—Cognitive operations in anglicized Moroccan fragrance brand names 153

María Enriqueta Cortés de los Ríos
Fatima Azzahraa El Yamlahi

Chapter 10.—A cognitive-pragmatic account of axiological neutrality in ironic constructions 167

Inés Lozano Palacio

Chapter 11.—Revisiting understatement: A cognitive-linguistic approach. 179

Francisco José Ruiz de Mendoza Ibáñez
María Sandra Peña-Cervel

III
KNOWLEDGE REPRESENTATION FOR SYNTAX AND LANGUAGE

Chapter 12.—Encoding reference phrases into AVMs in the Formalised Lexical-Constructional Grammar 195

Emmanuel Hernández Hernández

Chapter 13.—Interrogatives in the Formalized Lexical-Constructional Model (FL-CxG): The case of polar questions. 211

María Auxiliadora Martín-Díaz
Marta González-Orta

Chapter 14.—Feature-based representations of the ‘spray/load’ alternation: A Formalised Lexical-Constructional Grammar account 227

Carolina Rodríguez-Juárez

Acknowledgements

This publication is part of the I+D+i PID2020-112827GB-I00 project, funded by MCIN/AEI/10.13039/501100011033.

MINISTERIO
DE CIENCIA
E INNOVACIÓN

AGENCIA
ESTATAL DE
INVESTIGACIÓN

Chapter 4
Social media detection of texts on the social problem of violence against women within the multimodal intelligent system ALLEGRO

Ángela Alameda Hernández
https://orcid.org/0000-0002-8569-3533
Universidad de Granada

Abstract

In our modern society, violence against women is one of the most widespread and prevalent social problems affecting the integrity of this segment of society. In the era of social media, their users have become effective providers of information on this or other issues concerning society through the content they share publicly. Hence, detection and analysis of texts posted on the social media can have a practical impact to identify and take appropriate action against this situation. In this line, this chapter illustrates the process for the automatic detection of problems of violence against women within ALLEGRO, a multimodal intelligent system. More specifically, this work shows the elaboration of the conceptual schema that represents this social problem, as well as the compilation of a corpus of Twitter (now X) microtexts that allows the researcher to evaluate the effectiveness of the system in filtering appropriate tweets.

Keywords: social media, event detection, social problems, violence against women, ALLEGRO.

I. Introduction

Quality of life is not only related to economic prosperity but includes other factors beyond material goods. Among these intangible factors, physical safety is an essential element in ensuring people's well-being and includes both crime rates and the subjective perception of insecurity. A salient problem that affects physical safety in our current society is gender-based violence, and more particularly violent acts whose victims are women. Besides, in the age of digital social networks, their users have become effective providers of information through the content they post online. This information is published daily and on a massive scale, and it includes content related to events that may affect citizens' physical safety and, more particularly, the above-mentioned integrity of women. Thus, developing computational models capable of processing this huge amount of information would allow us to optimize the early warning systems we already have, as well as to create applications to improve the lives of citizens. Since the main challenge is to ensure that the detection of these problematic events is accurate and instantaneous, this chapter proposes an approach to the problems afflicting society from a sociolinguistic and a natural language processing (NLP) perspective.

On the one hand, this work is based on the study of primary sociological research sources that provided the factors that give rise to complaints, discomfort, or suffering among citizens, showing the trends in areas of concern emerging in our society. From the study of these sources, we can create an ontological taxonomy of recurring problems that can then be used to build an intelligent system for detecting these or related problems in the social media. In addition, this chapter shows the elaboration of linguistic schemas that represent those social problems, as well as the collection of a corpus of microtexts from the Twitter social network (now X) that allows the researcher to evaluate the effectiveness of such schemas in filtering appropriate tweets.

On the other hand, the present research illustrates the process for the automatic detection of physical-safety problems, particularly gender-based violence problems, within the multimodal intelligent system ALLEGRO, so that the taxonomy of problems and schemas that represent those problems are integrated into this system. ALLEGRO (**A**daptive mu**L**ti-domain socia**L**-media s**E**nsin**G** f**R**amew**O**rk) is a system, currently under development (http://allegro.ucam.edu), for the multimodal analysis of social media data, aimed at the detection of problems and adverse events in real time. ALLEGRO is made up of different modules for text, image, and sound processing, but, given that the scope of this chapter is NLP, it will only focus on one of them, namely, the DIAPASON module, as it is the one designed for linguistic processing and hosting lexico-conceptual schemas.

The remainder of this chapter is organized as follows. The following section reviews the literature on the detection of problematic events and the role of social media in detection systems. The third section analyses some of the most representative primary sources of information related to the area of public safety and gender-based violence specifically. The fourth section introduces the ALLEGRO system focusing on the DIAPASON module. It then illustrates the procedure for the creation of problem schemas and the elaboration of a corpus of microtexts from the Twitter/X social network.

II. Event detection and the social networks

Event detection is a relatively novel research area aimed at locating any kind of events, including social problems, through the extraction of the information about such events which is present in a given text, be it from news media or other platforms. This field has developed notably in the last decades (Yang *et al.*, 1998; Atefeh & Khreich, 2015; Afyouni *et al.*, 2022). The task is nonetheless challenging since the speed and the impressive amount of information that is spread daily makes manual analysis inoperative and has led researchers to the development of algorithms that, through the processing of written texts, can automatically detect events from the data. In this sense, event detection is a preprocessing step for many NLP tasks. In addition, since nowadays a large part of public incidents information is disseminated through social media networks, social media data have become a key source of analysis in this field (Panagiotou *et al.*, 2016). The analysis of events generated by social media users is performed using deep learning algorithms, both supervised and unsupervised (Balaji *et al.*, 2021), as well as big data analytics (Abkenar *et al.*, 2021).

One of the possible applications of this discipline is the creation of tools for the so-called 'smart cities', such as those aimed at improving different areas, including traffic management, security, or the functioning of public bodies, among others. NLP techniques, then, contribute decisively to the identification of events and, by analyzing the situations, they can be helpful for organizations to take the necessary action. This way, they have become a fundamental step in the process of creating smarter cities that are friendlier to their inhabitants (Cai, 2021). Also related to event detection, there is a further discipline known as crowdsensing or mobile crowdsensing (Ganti *et al.*, 2011), which is mainly focused on the analysis of data collected by physical sensors, such as those included in mobile devices (Xu *et al.*, 2018; Boubiche *et al.*, 2019). In this chapter, the concept of crowdsensing is conceived in a broad sense to include not only the monitoring of physical devices (mobile or fixed), but also to suggest the monitoring of social events from a more sociological perspective in which people, as generators of content in the social media networks, play a much more active and conscious role.

When a social problem, such as a gender-based violence incident or concern, occurs in the real world, one of the consequences is that it may motivate people to discuss it in social media and it will, then, be reflected in network activity. In the case of the Twitter/X platform, it means that a higher number of tweets will be posted on a specific time spam. Hence, the detection of a higher volume of tweets about a certain social problem is valuable information reflecting current problems affecting society. Indeed, data generated by social media users, known as UGC (user generated content), is especially valued because it is authentic and reflects aspects of real-world society, as opposed to institution-generated or company-generated content that respond to other goals such as educational, advertising, or journalistic. In addition, social media are heterogeneous sources of data since they include written text, as well as image and sound. Data fusion is the process of hybrid integration capable of linking these three types of data (Lau *et al.*, 2019). The computational system adopted for the processing of this complex and heterogenous social media data in the present research is ALLEGRO, which was conceived to integrate the three types of data. However, in this chapter the focus will be on the detection of social problems in the written texts disseminated in the social media. A specific characteristic of social media text is that it is not usually well-formed high-quality text, but rather has unique textual characteristics that include abbreviations, conversational nature, sentence fragments, use of nonstandard language, and misspellings. Many attempts have been made to make possible the challenging task of applying NLP techniques to this type of language (Liu *et al.*, 2012). Nevertheless, the application of event detection to social media data together with a focus on current social problems affecting people can have a remarkable and practical social impact.

III. Violence against women as a social problem. A review of sociological sources

Scrutiny of bibliographical sources on sociological research lays the groundwork for the whole process of computational modelling of social problems, as these sources provide an empirical view of the issues that affect citizens. In the framework of the ALLEGRO project, Periñán-Pascual (2023) establishes a general approach to the primary sources that address issues related to social welfare and people's happiness. Similarly, in the present chapter, the modelling of problems related to physical safety –as a broad concept where gender-based violence is included– has been done on the basis of scholarly literature, reports, statistical indexes, and specialized publications by relevant institutions both national and international, on the physical security of citizens to identify those areas on which there is a community reaction, and consequently, become social problems. As suggested in Eitzen *et al.*'s (2014) classification, physical-safety problems exemplify one of the two basic types of social problems, namely, those that refer to acts that violate the norms and values of a society. The concept of physical safety here refers to the state of being protected from any situation –such as crime or delinquency– that puts personal physical integrity at risk. Hence, situations of insecurity caused by accidents, risks of various kinds, or adverse natural phenomena will therefore be excluded from the interest of the present research. Although the last decades have seen a decline in crime rates (Tonry, 2014), a review of sources shows that social problems related to lawbreaking and offending are a focus of concern in sociological studies that present the broad spectrum of social problems.

Among these deliberate situatins that affect personal physical safety, our modern society suffers from violence against women to a considerable degree. Indeed, according to the UNWomen organization (2022), "violence against women is one of the most prevalent human rights violations in the world". It has a massive social impact, as it knows no economic, national, or social group boundaries. Gender-based violence is an umbrella term that covers all forms of harmful acts that are based on unequal power relations and are directed to individuals based on their gender roles (Mooney *et al.*, 2021). In specialized sociological research, as reviewed by Shepherd (2008), gendered violence is differentiated from violence against women in the sense that the former refers to more general unequal power relations, while the latter focuses on the identity of women as victims and men as perpetrators of violence. However, even if in this sense, gender-based violence includes a wider range of gender groups as victims, it disproportionally affects women and girls (UN Population Fund, 2022) and it is for that reason that this term is frequently used as a synonym for 'violence against women', and also intimately related to other terms such as 'domestic violence' or 'intimate partner violence', which are all ultimately rooted in gender inequality. In addition, violence against women does not only include acts that result in physical or psychological harm to women, but also threats of such acts. Hence, understanding this social problem should include both objective incidents, in the form, for example, of crime rates, as well as the subjective feeling of fear of being a potential victim of such acts.

Numerous national and international institutions include references to women's physical safety problems in their reports on crime statistics and general indicators of society's wellbeing and quality of life. To start with, in Spain, INE (*Instituto Nacional de Estadística*, 2022) considers objective and quantifiable indicators such as the number of criminal actions committed, as well as subjective or personal assessment aspects, such as citizen perception of security based on surveys. On the one hand, the crime indicators make an explicit reference to crimes against sexual freedom and indemnity. On the other hand, the subjective quality indicators refer to the conditions of the environment in which people live and make a specific reference to the perception of safety by the population when walking alone at night, which is again directly related to the female population. For its part, in the United Kingdom, the Office for National Statistics (ONS) in its 2022 bulletin on quality of life in this country includes the incidence of crime in relation to the safety conditions of the area where individuals live. In addition, its 2022 research update on violence against women and girls deals with the crime types that violence against women covers (homicide, sexual assault, stalking, and sexual exploitation, among others) for the statistical analysis of the incidence of this problem. It contains data from the Crime Survey for England and Wales (CSEW), which is especially valuable, as it collects information on crimes committed but not reported to the police. Finally, at the European level, the European Statistical Office (Eurostat) in its 2017 report (Eurostat, 2017) highlights that the central factor that puts physical security at risk is crime and violence whether it materializes or is only feared, and whether it is suffered to an extreme degree that could lead to death or in a milder form. In its July 2021 update (Eurostat, 2021), it shows that both the number of crimes, namely homicides, and the perception of violence have decreased in Europe between 2010 and 2019. However, no specific references to violence against women or women's feeling of physical safety are included.

Gender-based violence implies a violation of basic human rights and as such it is penalized internationally. The United Nations has called on all countries to take action to prevent it and to prosecute those who perpetrate such violence. There are several international treaties and conventions that address violence against women, including the Convention on the Elimination of All Forms of Discrimination Against Women and the Declaration on the Elimination of Violence Against Women. Many countries have also enacted laws to address violence against women and to provide legal protections and remedies for victims. However, gender-based violence has been historically sustained by a custom of silence and absence of denunciation of the situation by the victims, generally motivated by fear and a sense of powerlessness. In this sense, the new media of mass communication, such as the social media networks, have become potent loudspeakers for the disclosure and display of such practices. When detected and analyzed, the public exposure of this social problem can help governments and other institutions to take appropriate action.

IV. The integration of the problem of violence against women into the ALLEGRO system

As already introduced, the ALLEGRO system has been developed for the automatic detection of social problems in the social media. The main novelty of this system compared to other existing ones lies in the heterogeneous use of knowledge engineering techniques, natural language processing, and deep machine learning using convolutional neural networks. Likewise, another characteristic of this proposal lies in its granularity, since, unlike other models, ALLEGRO is designed to detect events of a very diverse nature by means of universal ontological parameters, which allows it to specifically differentiate issues such as economic problems, physical violence, harassment, ageing, marital conflicts, etc. In addition, although ALLEGRO integrates three modules (SOUND, ADAGIO, and DIAPASON) for the processing of the three types of data (sound, image, and text, respectively) disseminated in the social media, this chapter focuses on the third of these modules, since the goal is the detection of social problems only in written UGC units.

Thus, DIAPASON is the text analysis module and can be considered as the starting point for the whole process of social media data analysis, as knowledge will then be complemented, corroborated, or rebutted with sound or visual data. In terms of linguistic processing, its main contribution is the use of abstract schemas for the formal representation of citizens' problems regardless of the language they use to express them. This is made possible because this module is conceived as an ontology that stores the schemas for specific social problems in a taxonomic hierarchy. The hierarchy includes 3 main levels of specification: realm, dimension, and domain, where each specific social problem type must be inserted. At the top level of this hierarchy, there are two realms: physical and social. Since the former deals with more tangible aspects of the environment where individuals live, the social problem, which is the object of study of the present research, relates to the second realm. This social realm is itself divided into the Living, Economy, and Governance dimensions (Periñán-Pascual, 2024). The social problem concerned with violence against women belongs to the Living dimension, since this dimension covers a variety of domains that deal with those aspects that affect or determine the quality of life of people in society. Furthermore, this hierarchical ontological organization illustrates the thematic scope of DIAPASON and its potential to be populated with social problems of any kind.

It should also be noted that, to expand its detection and processing capacity, the ontology is designed to be flexible, i.e. a problem type can be related to more than one domain. Hence, the analysis of a variety of sociological sources and quality of life reports, as presented in section III of this chapter, has led us to identify violence against women as a social problem that should be integrated in more than one of the eleven domains in the Living dimension of DIAPASON. First, it populates the CrimeJustice domain since, gender-based violence is a violation of human rights and as such it is listed as a crime type and penalized internationally. Second, it populates the Inequality domain because it is a form of gender discrimination which is based on unequal power relations.

However, violence against women also relates to other domains such as Health, due to the mental health issues it causes on women who suffer it –psychological harm as well as problems in their reproductive health, among others–, and the SubjectiveWellBeing domain, because of the feeling of fear of being a potential victim of violent acts. As stated above, this complexity poses no problem, since the DIAPASON ontology permits the integration of a specific problem type into more than one dimension and domain, showing the capability of this system to reproduce and represent the complexity of our modern society and the concept of 'smart cities', in particular.

1. Conceptualization of the problem type 'violence against women'

To process information about each problem type, the DIAPASON module stores them in the form of problem schemas, i.e. formal representations that define the problem. Consequently, for the computational system to effectively use knowledge about the problem type of violence against women, a basic definition of the problem in natural language was elaborated as a necessary previous step for the elaboration of the problem schema. The definition basically consists in a precise, concise, and clear statement in natural language about the problem, including the basic terms that describe it. Violence against women is defined as "violent physical or psychological acts directed at women or girls". This definition is the basis for the elaboration of the schema. Additionally, the problem is given a formal name, namely, ViolenceAgainstWomen. As this specific case illustrates, the formal name is a list of running characters without spaces, using capital letters for the initial of each word in the name. Moreover, the name can contain only content words, hence functional words that would naturally be used to designate the problem are omitted.

Formally, schemas are composed of one or more basic units extracted from the WordNet database (Princeton University, 2010), which assigns a numerical code to a concept expressed in English. Each concept is linked to a set of synsets, or near synonyms, which increases the expressive capacity of the schemas in DIAPASON. From a procedural point of view, and to assist in the task of schema modelling, the DIAPASON editing interface contains a specific section for the *in situ* search of WordNet synsets, both in Spanish and English. Additionally, using WordNet synsets allows for multilingual problem processing, independently of the language in which the detection task is carried out. The choice of a formalism based on a natural language such as English makes it easier for knowledge engineers to create the schemas. Example (1) shows the schema designed for the problem type ViolenceAgainstWomen:

(1) ((violence-100965404 | maltreatment-100419908 | threat-106733476) & (womanhood-108477634 | girl-109992837))

Three synsets head the semantic core of the problem and are connected with the inclusive logical operator '|', which indicates disjunction and coordination (similar to 'and/or'). They refer to physical violence and maltreatment, but also to 'threats', that is, the declaration of an intention to inflict harm. Analysis of scholarly sources motivated the selection of these synsets that refer to the physical acts of aggression themselves, as well as to the declaration of their intention, which equally produces mental harm and health problems to women and are, thus, comprised in the understanding of the problem of violence against women. In addition, it should be noted that the selected synsets are semantically linked to other concepts, such as 'abuse' or 'force', that would also be used by the non-specialized users of the language to refer to acts of violence against women. The second expression, containing the synsets for 'women' and 'girls', is logically linked to the previous expression with the '&' operator, since concepts from both expressions formulate the conceptual definition of the problem.

2. **Compilation of a corpus of microtexts on violence against women**

To evaluate the validity of the schema designed to represent the problem of violence against women (ViolenceAgainstWomen) to filter and detect appropriate data from the social media, a subcorpus of microtexts about this specific problem was compiled from the Twitter/X social network. It is a subcorpus in the sense that, in the framework of the ALLEGRO project, it will be one of the various subcorpora that will make up a general corpus on social problems belonging to the different domains in which the ontology in the DIAPASON module is structured.

To access the Twitter API, a specific application was used, namely, Twitter Searcher, which has been specifically developed for the purpose of this project (see Periñán-Pascual, 2024). This application automatically retrieved a collection of tweets based on specific query parameters, which included:

a. Search words: This basic query was made both in English and Spanish to build a subcorpus for each of the two languages. Based on scholarly evidence that the expressions 'gender-based violence', 'domestic violence', and 'violence against women' are used by the general language user interchangeably, these three terms were used as search words for relevant tweets. This poses no problem as to the accuracy of the corpus to actually comprise the problem of violence against women specifically since, at a later stage, the researcher manually filters the tweets retrieved by the application, following a series of criteria, which will be presented in the following paragraphs. Hence, these words together with their Spanish equivalents (*violencia mujer*, *violencia de género*, *violencia doméstica*) were used in the different queries. Other words or combinations of words included in the DIAPASON formal schema were also used, such as 'maltreatment', 'threat of harm', or 'threat of violence'. The word 'maltreatment' (and its Spanish counterpart *maltrato*) in combination with 'women' were particularly productive for the retrieval of tweets dealing with the problem of violence against women, since these are ordinary words used by the general non-specialized population.

b. Language: queries were made for tweets in either English or Spanish. Since the schema for violence against women in the ontological module is a conceptual representation and, hence, it is language independent, the subcorpus under construction intends to assess its effectiveness to detect tweets on this social problem in any of the two languages.

c. Date: Tweets were retrieved periodically for approximately two months, dating from 20 February to 30 April 2023. This time span coincided with the celebration of the International Women's Day (8 March), which put the problem of violence against women in the focus of attention of society, thus, having a reflection in higher social media activity.

d. Number: For each query, 100 tweets were retrieved.

The vast number of tweets retrieved by the Twitter Searcher app dealt with violence against women in different ways. Hence, a closer manual analysis of them showed that further filtering was needed according to a series of criteria. These criteria are based on the scholarly definition of 'social problem' itself, which, far from being univocal, lacks consensus and is not free from discussion. Most recent theories describe social problems as social situations that are felt inappropriate or dangerous by a segment of society (Eitzen *et al.*, 2014). According to this subjectivist approach, a social problem is the belief that a condition that affects some members of society is felt by that same society as threatening to their lives and the quality of their lives and, consequently, a remedy should be found. Thus, for a social problem to be identified as such, the key element is not the objective social condition which members of society can see or learn about, but primarily the subjective interpretation of that condition, that is, the belief that that situation is pernicious and deteriorates the quality of life of at least a segment of society. Additionally, this belief does not become a social problem until it is shared by members of society. In that sense, through social network activity people negatively react to certain conditions making their subjective perception public and, consequently, shared through claims making. Accordingly, a social problem is understood as a specific type of community problem. In this sense, the presence of a high number of tweets in the Twitter/X platform that reflect users' reactions to and subjective perceptions of situations of violence against women is an indicator of this belief to be a social problem.

Stemming from this definition, UGC units to be included in the corpus must present the social problem as contemporary, i.e. as a current or present belief. This does not exclude the possibility that its existence may stretch from the past and/or into the future. However, since what society considers a social problem may change throughout history (Mooney *et al.*, 2021), there must be a time framework for the identification of a specific social problem. The framework for the creation of the present corpus is restricted to tweets that refer to problems for our society nowadays. Hence, the selection process excludes tweets that, although posted within the selected time span, deal with problems that have affected society in the past, even if the condition that motivated the problem may still exists, but there is not a shared belief that perceives that as a problem at present. Likewise, hypothetical or imagined problems that could affect society in the future are not suitable for the corpus. As a rule, messages that contain grammatical structures such as the past tense form of verbs, the modal 'will' with a future meaning in English, or the subjunctive verb form in Spanish when referring to a social problem are discarded as potential elements for the corpus. To illustrate this, example (2) presents the problem of violence against women as a contemporary social concern:

> (2) Why is violence against women suddenly ok in the so-called free world – what devilry is this? (10/4/2023)

Furthermore, for the UGC unit to be included in the corpus, it must be self-contained, i.e. the tweet must be able to be understood as a social problem independently of the cultural background knowledge or cultural identity of the reader. Any proficient speaker of the language in which the tweet is written must be able to understand the message and identify it as dealing with the social problem, even if they do not share the same cultural experience as the writer. This is particularly relevant in connection with references to names or the use of abbreviations that may refer to people, places, or institutions that are only locally or nationally known, particularly only temporarily known. Example (3) from the Spanish corpus illustrates this point:

> (3) En esta era estamos viviendo el mundo al revés, si RC representa 8M que baje Dios y lo vea, el 8M se lo carga. ["In this era, we live in an upside-down world, if RC represents 8M may God come down and see it, she will ruin 8M"]. (5/3/2023)

In this example, the initialism RC stands for the proper name of a Spanish celebrity who, at the time was regularly present on TV shows to denounce violence from her husband. Her person and her situation are only nationally and temporarily recognized. Those tweets are discarded to make the corpus as universally valid as possible. Other acronyms or initialisms that have become elements of the language, particularly slang and youth textspeak, are acceptable (e.g. lol, btw, 4U) as identifying features of a specific register. Likewise, those that belong to the terminology of the specialized field the social problem belongs to are accepted. Some illustrative examples are IWD (International Women's Day) or 8M (Spanish for March 8, the date when IWD is celebrated), which are regularly used in media content that deals with the problem of violence against women.

The filtering also discards UGC items that, although mentioning some problem theme, do not have a problem content. Hence, in the process for the creation of the corpus, it is important to differentiate and filter the tweets that deal with the subjective belief as a problem affecting society and discard those tweets that are merely referential and deal with the social problem as a topic. Therefore, tweets that inform on conferences, meetings, or that report on the existence of institutions or documents related to the social problem are not considered suitable for the corpus, since violence against women is only present as problem theme. Examples (4), (5), and (6) are illustrative cases in point:

> (4) Survivors of domestic violence, sexual assault and stalking under the Violence Against Women Act are eligible to receive additional resources from the U.S. Department

of Housing and Urban Development (HUD). Check out this article to learn more. (10/04/2023)

(5) How does domestic violence show up in our lives? How can we best prevent it from happening & protect victims of it? @profwschiller & later Prof April Zeoli joins #DetroitToday to discuss Comment w/#DetroitToday, call 3135771019, listen 101.9 FM, stream @wdet.org. (14/04/2023)

(6) Hiatus House offers a safe, caring environment for women and children who have experienced domestic violence. If you need help, call our 24/7 crisis line at 519-252-7781 #heretohelp #aplacetoheal, (14/04/2023)

On the contrary, the following tweets are illustrative examples of UCG units suitable for the subcorpus on the social problem of violence against women:

(7) Que impotencia siento por muchas razones… Tengo mucha rabia... Nada justifica un mínimo acto de maltrato hacia una mujer. [“What helplessness I feel for so many reasI.... I am very anI... Nothing justifies the slightest act of mistreatment towards a woman”]. (06/03/2023)

(8) I hate this tradition. It’s rooted in misogyny and normalizes violence against women. When it comes to gender equality, this country is so fucking backwards. (10/04/2023)

(9) Women and girls deserve to live free of violence. It’s time to step up, to act and to end violence against women & girls. (10/04/2023)

These tweets illustrate problem content, as they reproduce media users’ subjective perceptions and negative reactions to violent acts against women. Eventually, after the refined compilation of tweets about the problem of violence against women, the final goal is to create a gold standard corpus that can be used to evaluate the accuracy of the ALLEGRO system in the automatic and machine-driven detection of social problems in the social networks. In the NLP context, a gold standard corpus is a manually annotated collection of text (Wissler *et al.*, 2014). For the subsequent annotation of this corpus, the INCEpTION platform (Klie *et al.*, 2018) will be used, but its description and actual annotation process fall outside the scope of the present chapter.

V. Conclusion

The application of automatic event detection to social media data, together with a focus on the analysis of current social problems affecting citizens, can have a remarkable and practical social impact. To this end, this chapter has shown a practical application within the framework of the ALLEGRO project and with a focus on the social problem of violence against women. Based on the study of sociological sources, formal schemas that represent this social problem were elaborated to populate the DIAPASON module, responsible for linguistic processing. The chapter has finally described the process and the criteria for the collection of a corpus of UGC units from the Twitter/X social platform for the evaluation of the automatic detection of the social problem of violence against women by the ALLEGRO system.

VI. References

Abkenar, Sepideh Bazzaz, Mostafa Haghi Kashani, Ebrahim Mahdipour & Seyed Mahdi Jameii (2021) Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics* 57: 101517.

Afyouni, Imad, Zaher Al Aghbari & Reshma Abdul Razack (2022) Multi-feature, multimodal, and multi-source social event detection: A comprehensive survey. *Information Fusion* 79: 279–308.

Appio, Francesco Paolo, Marcos Lima & Sotirios Paroutis (2019) Understanding smart cities: Innovation ecosystems, technological advancements, and societal challenges. *Technological Forecasting & Social Change* 142: 1–14.

Balaji, TK, Chandra Sekhara Rao Annavarapu & Annushree Bablani (2021) Machine learning algorithms for social media analysis: A survey. *Computer Science Review* 40: 100395.

Boubiche, Djallel Eddine, Muhammad Imran, Aneela Maqsood & Muhammad Shoaib (2019) Mobile crowd sensing: Taxonomy, applications, challenges, and solutions. *Computers in Human Behavior* 101: 352–370.

Cai, Meng (2021) Natural language processing for urban research: A systematic review. *Heliyon* 7(3): e06322.

Eitzen, D. Stanley, Maxine Baca Zinn & Kelly Eitzen Smith (2014) *Social Problems*. 13th ed. Allyn and Bacon.

Eurostat (2017) *Final Report of the Expert Group on Quality of Life Indicators*. Publications Office of the European Union.

— (2021) *Quality of Life Indicators: Economic Security and Physical Safety*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Quality_of_life_indicators_-_economic_security_and_physical_safety.

Ganti, Raghu K, Fan Ye & Hui Lei (2011) Mobile crowdsensing: Current state and future challenges. *IEEE Communications Magazine* 49(11): 32–39.

Instituto Nacional de Estadística. INE. https://ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735573113.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho & Iryna Gurevych (2018) The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics* (COLING 2018). Association for Computational Linguistics, 5–9.

Lau, Billy Pik Lik, Sumudu Hasala Marakkalage, Yuren Zhou, Naveed Ul Hassan, Chau Yuen, Meng Zhang & U-Xuan Tan (2019) A survey of data fusion in smart city applications. *Information Fusion* 52: 357–374.

Liu, Fei, Fuliang Weng & Xiao Jiang (2012) A broad-coverage normalization system for social media language. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1035–1044.

Mooney, Linda A., Molly Clever & Marieke Van Willigen (2021). *Understanding Social Problems*. Cengage learning.

Princeton University (2010) About WordNet. *WordNet*. Princeton University.

United Nations Population Fund (2022) Gender-based violence. https://www.unfpa.org/gender-based-violence.

Office for National Statistics (2022) Crime survey for England and Wales. https://www.crimesurvey.co.uk/en/index.html.

— (2022) Quality of life in the UK: August 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/qualityoflifeintheuk/august2022.

— (2022) Violence against women and girls: research update November 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/violenceagainstwomenandgirls/researchupdatenovember2022.

Panagiotou, Nikolaos, Ioannis Katakis & Dimitros Gunopulos (2016) Detecting events in online social networks: Definitions, trends and challenges. In Michaelis, Stefan; Nico Piatkowski & Marco Stolpe (eds.) *Solving Large Scale Learning Tasks. Challenges and Algorithms*. Springer Cham, 42–84.

Periñán-Pascual, Carlos (2023) Modelización de las quejas de los ciudadanos como artefactos digitales culturales: DIAPASON. In Olmo-Cazevieille, Francoise (ed.). *Investigación Lingüística en Entornos Digitales*. Tirant Lo Blanch.

— (2024) Exploring problems through social media: The case of beach quality. In Jiménez-Briones, Rocío & Avelino Corral Esteban (eds.) *Approaches to Knowledge Representation and Language*. Comares.

Tonry, Michael (2014) Why crime rates are falling throughout the western world. *Crime and Justice* 43: 1–63.

Shepherd, Laura (2008) *Gender, Violence, and Security: Discourse as Practice*. Zed books.

UNWomen (2022) Types of violence against women and girls. https://www.unwomen.org/en/what-we-do/ending-violence-against-women/faqs/types-of-violence.

Wissler, Lars, Mohammed Almashraee, Dagmar Monett Díaz & Adrian Paschke (2014) The gold standard in corpus annotation. *IEEE GSC*. IEEE.

Xu, Zheng, Lin Mei, Kim-Kwang Raymond Choo, Zhihan Lv, Chuanping Hu, Xiangfeng Luo & Yunhuai Liu (2018) Mobile crowd sensing of human-like intelligence using social sensors: A survey. *Neurocomputing* 279: 3–10.

Yang, Yiming, Tom Pierce & Jaime Carbonell (1998) A study of retrospective and on-line event detection. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computational Linguistics, 28–36.

VII. Acknowledgements

Financial support for this research has been provided under grant PID2020-112827GB-I00, funded by MCIN/AEI/10.13039/501100011033.