Expert Systems With Applications 208 (2022) 118063

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise

J. Angel Diaz-Garcia *, M. Dolores Ruiz, Maria J. Martin-Bautista

Department of Computer Science and Artificial Intelligence, University of Granada, Daniel Saucedo Aranda, s/n, 18014 Granada, Spain

ARTICLE INFO

Keywords:
Social media mining
Pre-processing
Credibility
Word embeddings

ABSTRACT

Social networks have taken an irreplaceable role in our lives. They are used daily by millions of people to communicate and inform themselves. This success has also led to a lot of irrelevant content and even misinformation on social media. In this paper, we propose a user-centred framework to reduce the amount of irrelevant content in social networks to support further stages of data mining processes. The system also helps in the reduction of misinformation in social networks, since it selects credible and reputable users. The system is based on the belief that if a user is credible then their content will be credible. Our proposal uses word embeddings in a first stage, to create a set of interesting users according to their expertise. After that, in a later stage, it employs social network metrics to further narrow down the relevant users according to their credibility in the network. To validate the framework, it has been tested with two real Big Data problems on Twitter. One related to COVID-19 tweets and the other to last United States elections on 3rd November. Both are problems in which finding relevant content may be difficult due to the large amount of data published during the last years. The proposed framework, called NOFACE, reduces the number of irrelevant users posting about the topic, taking only those that have a higher credibility, and thus giving interesting information about the selected topic. This entails a reduction of irrelevant information, mitigating therefore the presence of misinformation on a posterior data mining method application, improving the obtained results, as it is illustrated in the mentioned two topics using clustering, association rules and LDA techniques.

1. Introduction

Social networks have become an essential part of our lives. They are great sources of information, used daily by thousands of people to explore news and share their opinions about them. This great success has also led to the increasing spread of irrelevant information, hoaxes or misinformation, even interfering in electoral processes (Allcott & Gentzkow, 2017). Twitter, is one of the most successful social networks today, and undoubtedly the most used social network to share and comment on news around the world. Its character is mainly public so anyone can see that someone else is tweeting about a certain topic. This has led to Twitter gain on a relevant role, for example, to obtain relevant information in real time about events and disasters, but it has also made it a target for those who want to spread misinformation. But, what are relevant information and misinformation?

In our field of application, relevant information is understood as information that may contain valuable content in a certain context. For example, in a health topic, relevant information would be that issued by a doctor about prevention measures for a certain disease. That is, in the case of Twitter, for the proposed case of health, we will be interested in keeping those candidates (tweets) of relevance to medicine, discarding those samples (tweets) that are not related to this sector.

As for misinformation or disinformation, there are more and more papers that provide a description or new characteristics of this concept (Aswani, Kar, & Ilavarasan, 2019; Kar & Aswani, 2021). Specifically, in Aswani et al. (2019) they provide a very interesting vision of misinformation seen in different ways such as wilful misinformation, fictional discussions, and non-verifiable information or news. In any case, it is untruthful information that is disseminated for various purposes, such as to negatively influence political issues. In these cases where there is a clear intention to disseminate false content, we will speak of disinformation. Therefore, the difference between misinformation and disinformation lies in the intentionality or purpose. In the scope of our paper, we will focus on experience-driven misinformation reduction, since a person on Twitter may share false content, because of their beliefs without actually knowing whether that is false to a greater or lesser extent.

Being able to discern what is true or relevant in social networks and what is not, has taken up a great amount of literature in recent

* Corresponding author.
E-mail addresses: joseangeldiazg@ugr.es (J.A. Diaz-Garcia), mdruiz@decsai.ugr.es (M.D. Ruiz), mbautis@decsai.ugr.es (M.J. Martin-Bautista).

https://doi.org/10.1016/j.eswa.2022.118063
Received 25 October 2021; Received in revised form 30 June 2022; Accepted 3 July 2022
Available online 13 July 2022

years (Oehmichen et al., 2019; Shu, Sliva, Wang, Tang, & Liu, 2017), with artificial intelligence systems assuming a great importance in the process. The process of eliminating misinformation on social networks is, by its nature, closely linked to the processes of dimensionality reduction and instance selection, as both seek to eliminate data that is not interesting for subsequent data mining processes. This process can be guided by statistical methods such as features or instance selection algorithms (Olvera-López, Carrasco-Ochoa, Martínez-Trinidad, & Kittler, 2010), or by objective credibility data in the case of misinformation detection.

Our goal is to design a framework to address the problem of relevant content selection in social networks. With this objective, we seek to obtain smaller and more cohesive datasets on which to obtain better knowledge with subsequent data mining processes. From this main objective, the elimination of misinformation can be derived. The framework will select only those accounts with experience and credibility, which will therefore eliminate to some extent the possible misinformation present in the dataset.

In this paper, we propose a framework based on iterative filters, word embeddings, user authority and credibility, to reduce the irrelevant content of data and, at the same time, increase the confidence of retrieved information coming from social networks. With our framework, we achieve a more cohesive, clean and truthful dataset that can be used in subsequent processes with greater accuracy regarding the credibility of the source and the data. The system is based on the premise that if a user is credible, his or her content will also be credible. Therefore the proposed system will identify which users are credible and get their tweets according to a specific topic. The major contributions of the work to the state of the art are as follows:

- A new framework based on iterative filters, word embeddings and credibility is proposed for instance reduction in social media.
- A new method for filtering irrelevant information in social networks is proposed.
- A new algorithm is proposed for the selection of credible users in the social network Twitter based on the popularity and expertise of the user. To do this we focus on the user's biographies and process it with word embedding. As far as we know, this is the first work that applies word embeddings techniques to the biographies of the user on Twitter, using this to discern the user's expertise on a certain topic.
- The functionality and versatility of the proposed pre-processing system that can be used with a wide variety of data mining techniques, especially those sensitive to the amount of data. The framework has been tested with LDA, association rules and clustering techniques.

In order to validate the system, a set of experiments has been devised in which K-means Clustering (MacQueen et al., 1967), Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) and Apriori algorithm (Agrawal, Srikant, et al., 1994) are applied to large sets of data from Twitter, one related to the COVID-19, and another one concerning the United States elections. The experiments show the differences in the results when NOFACE (NOise Filtering According Credibility and Expertise) framework is applied or not. The results illustrate the reduction of content, processing efficiency, achieving an improvement of the algorithms used. This improvement will come in terms of a more manageable dataset and execution times.

The paper is organised as follows. In the following section, we study related work and different approaches for instance selection and dimensionality reduction according to credibility in social networks. In Section 3 we go into detail in the NOFACE framework describing each of the constituent modules. In Section 4, we explain the evaluation of the framework, which will be based on comparing the results of different data mining techniques, when NOFACE is applied or not. In Section 5, we discuss some of the challenges that this framework and others in the literature should face in the future. Finally, in Section 6, we examine the conclusions remarks and the future work of the research carried out.

2. Related work

The dimensionality reduction seeks to have clean data without noise, or loss of information, more understandable, and easily manageable and processable, something that highlights its relevance to Big Data problems. Dimensionality reduction usually focuses on the reduction of training variables (Solorio-Fernández, Carrasco-Ochoa, & Martínez-Trinidad, 2020), but there is also a branch dedicated to reduce or select the number of examples, which is where instance selection algorithms appear. By definition, the instance selection is framed within the techniques of data pre-processing and has been approached from multiple and different perspectives over the course of the years (Chandrashekar & Sahin, 2014; Olvera-López et al., 2010) . In problems related to social networks, where misinformation, noise and massive amounts of data are always part of the problem, these dimensionality reduction techniques are paramount. In this context, we focus on credibility-based dimensionality reduction techniques. In credibility analysis, the amount of data related to a problem is also reduced, but this reduction is guided by factors inherent to the social network where it is applied. These factors can be, for example, the followers or the engagement or the expertise of a user in a certain topic (Canini, Suh, & Pirolli, 2011). The analysis of credibility and the consequent reduction of noise and examples of the problem has been approached from different perspectives within Data Science and Artificial Intelligence. These perspectives depend on the techniques used and the granularity of the entity on which its credibility is studied. At this point, we have carried out a study of the state of the art of the credibility analysis in Twitter. We have classified the works according to their granularity in content (tweet or topic) level or user level.

Since the framework, by selecting relevant and credible content, can also help to eliminate misinformation, additionally we have conducted a literature review on this aspect.

2.1. Content level credibility

Castillo, Mendoza, and Poblete (2011) have addressed the problem of credibility on Twitter. Their research is one of the most comprehensive and attempts to classify contents (tweets) based on whether they are credible or not. To evaluate and create the model, they use a large number of indicators that are closely linked to the analysis that a human would do to study the credibility of a tweet, such as whether an account is verified, whether the user has enough followers or whether the tweet uses appropriate hashtags, to name some of the features taken into consideration by the classification system. Concerning the user, it also obtains information but in a very simplistic way: for example, if it has biography or if it is empty. At this point NOFACE goes one step further, analysing the biography completely and obtaining knowledge of it to guide the process of content filtering and dimensionality reduction.

Kang et al. offers in Kang, O'Donovan, and Höllerer (2012) three different ways to obtain a credibility rating. The first proposal analyses the social graph of Twitter, by means of ratios between concepts like retweets or number of followers. The second one focuses on content, and finally the third model is a hybrid model that takes into consideration graphs and content. Being the first model, the one based only on graphs, which works best.

Finally, there is also a credibility-oriented dimension to event-related content. Hassan (2018) uses text mining techniques on event-related tweets. The text mining techniques used are guided by the frequency of terms in different topics, and finally the algorithm is evaluated using different classifiers.

2.2. User level credibility

With regard to the analysis of user credibility, we find approaches such as those of Cognos or CredSaT. Cognos (Ghosh, Sharma, Ben-

evenuto, Ganguly, & Gummadi, 2012) offers a web solution for searching experts in a certain topic, for this, it uses Twitter lists. The lists on Twitter are user-managed lists, in which users add other users related to topics. Cognos exploits this potential, even improving the search for accounts in the native system of recommendation of Twitter. The CredSaT (Abu-Salih, Wongthongtham, Chan, & Zhu, 2019) system, is a Big Data solution that takes into consideration the content and the time stamp to create a ranking of expert and influential users in the social network. It also adds a semantic analysis layer with sentiment analysis on tweets and responses used to enrich the final corpus of experts.

Unlike these approaches, the NOFACE seeks to reduce the amount of data, that is, the aim is not to search for influential people but to guide content reduction of a social media dataset through expert users.

Finally, it is necessary to mention the papers proposed by Alrubaian, AL-Qurishi, Alrakhami, Hassan, and Alamri (2016) and Alrubaian, Al-Qurishi, Hassan, and Alamri (2018). These papers also deal with the analysis of credibility on Twitter in a very exhaustive way and similar to NOFACE through 3 modules. These modules deal with content credibility, reputation and expertise. However, NOFACE obtains the expertise according to the user's biography, instead of according to the content as made in Alrubaian et al. (2016) and Alrubaian et al. (2018). Our approach exploits the potential of biography in social networks such as Twitter, where it is very common to talk about professions. As far as we know, this is the first work that addresses and uses this option in addition to word embeddings, and with great results as we will see in future chapters. Additionally, our analysis of reputation is different to that of the above-mentioned papers focusing more on the engagement of user-generated content, which will give a value about how interesting is a user's content to his or her followers.

2.3. Misinformation detection

In the field of misinformation and fake news reduction, we find that supervised approaches are the most widespread. In Ozbay and Alatas (2020) Ozbay and Alatas, apply 23 different classification algorithms over a previously labelled fake news dataset coming from the political scene. With this same approach, we find the paper (Cordeiro, Pinheiro, Moreira, Carvalho, & Freire, 2019) in which, the authors apply again a battery of different classification methods that go from the traditional decision trees to the neural networks, all of them with great results. Also in the field of classification, but using bio-inspired algorithms, we find the paper (Batra, Jain, Tikkiwal, & Chakraborty, 2021). In this paper, the misinformation, or worthless information, comes in the form of email spam. The authors create a classifier based on K nearest neighbours and bio-inspired algorithms to obtain the instances that best represent the problem domain according to three different distance metrics.

If we look at the branch of deep learning, many papers have been used to detect fake news or rumours. One of the first is the one proposed in Ma et al. (2016). In this paper, they use an architecture based on three layers. The first layer uses as input the K most significant elements of the text based on the TF-IDF ratio to train a Recurrent Neural Network (RNN). Then it uses a Long Short-Term Memory (LSTM) layer to model the dependencies along with the text and in a final step it uses a layer based on a Gated Recurrent Unit (GRU). The model improves on the performance of other base models. Although the model gives good results, it is necessary to train the network, so it is necessary to re-train it for its use in another domain, because it is necessary to have labelled databases. This is something that does not need to be taken into account in systems based on data and user characteristics such as NOFACE and other models such as the one proposed in Castillo et al. (2011).

Also within the framework of deep learning there is a wide range of papers (Kaliyar, 2018; Molina-Solana, Amador Diaz Lopez, & Gomez, 2018; Monti, Frasca, Eynard, Mannion, & Bronstein, 2019). These papers have a similar focus. They use pre-labelled fake and non-fake databases to train classifiers based on neural networks. The proposal in Kumari, Ashok, Ghosal, and Ekbal (2021) is based on the use of concepts such as novelty and emotions to guide the detection of misinformation. Its foundation is the premise that this type of news and information tends to be emotionally charged to favour its diffusion. To do so, they use a combination of BERT, LSTM and feed-forward networks. In Nasir, Khan, and Varlamis (2021) authors also propose a combination of different neural network topologies. Specifically, they use Convolutional Neural Network (CNN) to obtain fake news features and applies in a later stage RNNs to store the sequential dependencies between terms. The output is then used for fake news classification.

Other papers use more novel approaches. For example, the work (Khoo, Chieu, Qian, & Jiang, 2020) use Twitter conversations generated around fake news to early detect the spread of fake news using neural networks. The work (Liu & Wu, 2020) uses the concatenation of user-related features and user-generated text to use them as input in a rumour classification layer applying word embeddings.

These previous approaches, although novel, also require training. Our model uses word embeddings in an unsupervised way, helping to reduce the amount of irrelevant data and, to some extent, mitigating the problem of false information. Importantly, NOFACE also uses a conjunction of user-based (number of favourites or retweets) and text-based (experience-related words present in the biographies) features. The potential of this feature fusion has been also highlighted in papers such as Liu and Wu (2020).

In summary, the major differences of the NOFACE framework compared to the solutions seen in the literature are:

- The main aim of NOFACE is the reduction of content coming from social networks considering the credibility and expertise of the publisher. To achieve this main objective, the reduction is guided by credibility, engagement and expertise analysis. These tasks are an intermediate stage of the main purpose, being, therefore one of the first methods of this kind.
- NOFACE offers a more restrictive cascade approach than other approaches where credibility, expertise or engagement is computed for all examples. NOFACE discards those that do not pass the first filter, the second filter and so on.
- NOFACE is, as far as we know, the first framework that applies word embeddings and text mining to the process of computing expertise through biographies.
- NOFACE offers an interpretable way to locate useful content in social networks and without having to train a classifier or neural network, so it can be used as a first stage of analysis on any dataset without the need to have a ground truth.

3. The NOFACE framework architecture

In this section we will go into detail in the NOFACE framework. In Fig. 1 we can see a general diagram of our framework NOFACE (NOise Filtering in social media According to Credibility and Expertise). The system is based on Twitter databases, on which the different modules are applied in cascade. The first module, focused on expertise, is based on a filter that uses word embeddings, concretely FastText (Bojanowski, Grave, Joulin, & Mikolov, 2016), to obtain those descriptions of users with greater expertise about a certain topic according to their profession. As far as we know, this is the first work that applies word embeddings on the descriptions instead of on the tweets. Once filtered by the users' expertise, a new filtering step is necessary to discard those profiles without interaction or credibility on the network. Afterwards, the next two modules, about the engagement and the credibility, are applied over the selected users by the first module. This new filter is based on the network metrics, such as the number of followers, if they publish valuable content, the favourites or the retweets. The final result is a very reduced set of data where tweets have been generated by people who not only have credibility on the net, but also have experience in the topic under analysis.

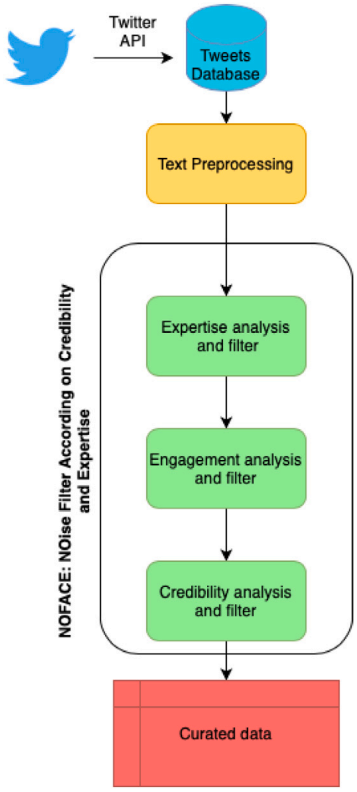

Fig. 1. NOFACE framework.

The central core of our framework is to be able to know which people talking about a certain topic on Twitter really have a relationship in terms of experience and credibility with the topic under study. Modelling credibility and expertise is an arduous task, approached by other papers in a quite exhaustive and efficient way (Alrubaian et al., 2016). Our framework is based on the premise that if a Twitter user is credible, his or her content is also credible and therefore the user can be used to guide the selection of the content.

Our framework is based on 3 modules plus a pre-processing module that are applied in cascade, that is, the users that do not pass the first filter, will not pass to the second one, which implies that we reduce computation times and generate a reduced and trustworthy solution. The framework, first applies a pre-processing method. It then computes the expertise through biographical analysis, then focuses on content quality (engagement) and finally filters based on the user's credibility on the topic. In Kumar, Kar, and Ilavarasan (2021) authors highlight the value of text mining applied in the field of credibility and fake information. They mention some approaches that corroborate our premise as valid. For example, in Barbado, Araque, and Iglesias (2019) authors detected a strong weight between several social features, such as lists or the number of followers, which are associated with more trustworthy content. The relation between misinformation and certain user-inherent components such as the longevity of the account, the presence of certain words in its content or the interaction generated with other users of social networks, has also been analysed in the state-of-the-art review in Wu, Morstatter, Carley, and Liu (2019). Therefore, we find feasible to create a system based on user features in conjunction with expertise to remove irrelevant information from social networks.

3.1. Pre-processing module

The pre-processing module uses common cleaning techniques in addition to others specific to the Twitter domain. All the techniques have been modularised in a Python function that allows to select the language of the text corpus, to guide sections like the elimination and detection of stop words. The pre-processing is applied to the user's biography and to the tweet content. The techniques used and their order of application are:

1. Twitter domain related cleaning. For this purpose we have eliminated URLs, hashtags, mentions, reserved words from Twitter (RT, FAV...), emojis and smileys.
2. Cleaning of the usual text mining domain, removing numbers, additional spaces and punctuation marks.
3. Turning the text into lowercase letters.
4. Detecting the tweet language, all those tweets using a non-recognised language or from a language other than the one desired by the user are eliminated.
5. The stop words of the language introduced by the user in the pre-processing function are removed.
6. Any empty tweets (composed of items eliminated in previous stages of pre-processing) are removed.
7. Tokenization of the biography and the tweet.

After this process, we have achieved that the raw content coming from Twitter, can be easily processable in later stages. Although NOFACE only uses data related to the users (biography, followers, etc.) it also cleans data related to the tweet content, to generate a final corpus useful for posterior data mining processes.

3.2. Expertise filter module

The first filter and therefore the main filter and greatest contribution to the state-of-the-art of this approach is the expertise filter. This filter must be able to locate and eliminate those users who are not really related to the topic. To do this, the filter will exploit the biographies of the users to the extreme. The Twitter code itself, also takes advantage of the biographies in its service *Who To Follow*, so we can conclude that using biographical content can be of a great interest. NOFACE will use the biographies to create a filter and through the use of word embeddings, it will exploit the potential of biographies in a very exhaustive way.

The vector space representation using word embeddings corresponds to the current state-of-the-art in Natural Language Processing (Levy & Goldberg, 2014; Liu, Liu, Chua, & Sun, 2015). The underlying technique is to represent all the words within a given vocabulary in vector space as vectors. With these vectors, operators such as addition or substraction can be applied, so that the words *king - man + woman* would result in the word *queen*.

Our expertise filter exploits this potential by using a representation in which each word is represented by a vector in the vector space. We can see the pseudo-code in Algorithm 1. Our algorithm uses the power of semantic relations between words to increase the search space in Twitter biographies. Many works have demonstrated the power of word embeddings to expand search queries. In Roy, Paul, Mitra, and Garain (2016) Roy et al. use the KNN algorithm on vector space generated by embeddings to obtain which terms are most similar to others and expand the search query. With a similar point of view we find the works (Diaz, Mitra, & Craswell, 2016; Kuzi, Shtok, & Kurland, 2016). In Diaz et al. (2016) Diaz et al. train locally embedding, namely GloVe (Pennington, Socher, & Manning, 2014) and Word2Vec (Mikolov, Chen, Corrado and Dean, 2013; Mikolov, Sutskever, Chen, Corrado and Dean, 2013), to improve search processes in information retrieval. In a very similar way but with Word2Vec+CBOW Kuzi et al. demonstrate in Kuzi et al. (2016) how document retrieval actually improves with this technique. In our algorithm, we will use this potential of word embeddings, not for retrieving documents, but for locating users that are experts on a topic.

Our algorithm works as follows. We introduce a list of words related to the topic under study, for example, about medicine, we can introduce words: *medical, doctor*. The algorithm will start to train a word embedding model on the biographies using a part of a data partition (10% of the entire dataset), and in the first iteration it will obtain the 5 most similar words to *medical* and *doctor* among the corpus itself. The algorithm will use these 12 words, (5 more similar to medical, 5 more similar to doctor, besides doctor and medical), to find users whose biographies contain any of these terms and start creating the list of experts and topic-related users. In the next iteration, we will already have 12 words to search for their 5 similar ones, and so on. This leads to an exponential growth of words linked by word embeddings to the domain of the problem. To prevent this and thus leading to a degeneration in meaning relative to the input words, the stopping condition of the algorithm is 3 iterations. This will enrich the space of words obtaining very related and linked to the topic words that will guide our filter.

In each iteration, the algorithm checks if any user id is already present in the expert list to avoid processing it again, since its words and content are already in the search corpus, thus avoiding additional processing. The output of the algorithm is a clean set of data in the form of a data frame ready to be processed in the following modules. It should be noted that at the end of the computation of the algorithm we will have a new column in the dataset, where we will see the words learned and used for filtering. In this way, a potential user can see in a readable and interpretable way what set of words the algorithm has used on each account to determine if it is a valuable account for analysis. The interpretability of the result of each step of the algorithm is therefore a value to be taken into account.

There are a multitude of models and representations for word embedding, so for fine-tuning our algorithm we have compared Word2Vec and FastText (Bojanowski et al., 2016), since they are the most widespread and relevant in terms of versatility and performance at present. The main difference between Word2Vec and FastText is that the latter decomposes each of the input words in the neural network into n-grams, for example for the word *matter*, and *n=3* we would have *{ma, mat, att, tte, ter, er }* and the final representation would be the sum of the vectors associated with each n-gram. This representation is very interesting to discern out-of-vocabulary words or words with low presence in the dataset. Word2Vec takes each word as a vector, therefore does not consume as much memory and resources as FastText (which for each word stores a vector per n-gram), although it is more sensitive to out-of-vocabulary words. For each of these embedding models, we have two representations, Skip-Gram and Continuous Bag of Words (CBOW). Skip-Gram tries to predict the context words surrounding the word, i.e. it predicts context based on a word. On the other hand, CBOW predicts a word based on the surrounding context words, i.e. it predicts a word based on the context. Regarding the embeddings parameters, it has been run with a window of 5 words, words with frequencies lower than 2 have been ignored and negative sampling of 10 has been done in the case of CBOW and a hierarchical softmax in the case of Skip-Gram. The dataset used to fine-tune the algorithm, is about COVID-19 (4.1.1) and is composed by 3 batches of 936.427, 1.062.900 and 1.319.912 tweets respectively (total 3.319.239 tweets). The results in the case of Word2Vec can be seen in Table 1. The results in the case of the experiments carried out with FastText are presented in Table 2.

In our problem, where we have few context words due to the fact that Twitter texts are not very large, this makes an important difference. It is easier for the algorithm to predict context words based on a single word (Skip-Gram), than to predict a single word based on several words (CBOW), since the search space and the window within each document (tweet) is very small. A priori it may seem that this does not influence, since Word2Vec+CBOW obtains great results, but if we compute the ratio of users found for each word, we can see how this value is 102 users per word on average in Word2Vec+CBOW, while this rises to 156 users per word in the case of Word2Vec+SkipGram. This ratio in the case of FastText+CBOW stands at 155, while in FastText+SkipGram the ratio reaches a value of 162. This leads us to conclude that the words

Algorithm 1: Expertise filter algorithm

```
Result: Dataframe with experts in the topic
# pre-processing, initialising the variables and data structures
cleaned-dataset=preprocess(dataset)
expert_set=[]
finaldataframe=pd.dataframe()
split cleaned-dataset into batches
for batch in batches do
    #we check if any user of the batch is already located as an
     expert
    if user_id in expert_set then
        # For experts, we add all their tweets to the final data
         frame and do not process
        finaldataframe.extend(batch[id=user_id])
    else
        # We process the rest of the content to locate new experts
        # get the biographies tokens
        tokenized_tweet = batch['biographies_clean']
        # train the word2vec model
        model=train_word2vec(tokenized_tweet)
        # create a list with the words of the list present in the
         model
        final_words=[]
        for word in expert_words do
            if word in model then
                final_words.append(word)
            end
        end
        # create a data frame with the 5 most similar words to
         each expert word
        for word in final_words do
            most_similar.extend(find_5_mosts_similar(model,
             word))
        end
        # extend the expert word list with the most similar words
         in the embedding
        expert_words.extend(most_similar)
        # We locate all users who have in their biographies any of
         the words
        new_experts =find_users(batch['biographies_clean'],
         expert_words )
        # Extend the final data frame with the tweets of the located
         experts
        finaldataframe=finaldataframe.extend(batch[user_id in
         new_experts])
        # Extend the expert set with the new experts
        expert_set.extend(new_experts)
    end
end
```

Table 1
Minimum and maximum value for each variable in the Word2Vec experiments.

	Word2Vec+CBOW	Word2Vec+Skip-Gram
Elapsed time	Min: 11 min 7 s Max: 13 min 36 s	Min: 13 min 1 s Max: 14 min 22 s
Words	Min: 209 Max: 228	Min: 45 Max: 64
Users located	Min: 21 390 Max: 23 377	Min: 7937 Max: 10 044
Final dataset size	Min: 31 347 Max: 34 107	Min: 12 281 Max: 15 239

Table 2
Minimum and maximum value for each variable in the FastText experiments.

	FastText + CBOW	FastText + Skip-Gram
Elapsed time	Min: 16 min Max: 17 min 15 s	Min: 18 min 20 s Max: 20 min 10 s
Words	Min: 50 Max: 79	Min: 111 Max: 125
Users located	Min: 10 975 Max: 12 312	Min: 18 128 Max: 20 306
Final dataset size	Min: 16 748 Max: 18 773	Min: 26 547 Max: 31 611

located by FastText have a higher representation in the dataset, as well as a higher relationship with the topic under study. Therefore, the best option for our algorithm will be to use FastText+SkipGram, although more time-consuming. This increase is also linked to a higher match value for the selected words and their relation to the topic, as well as a better user selection ratio.

3.3. Engagement filter module

The next filter concerns engagement. Engagement could be defined as the capacity of a user to generate useful content that is appreciated by other users of the social network. In other words, it is a measure of how good is the content a user publishes on a social network. In the specific case of Twitter, interaction is usually measured in terms of RTs (Retweets) and FAVs (Favourites). A RT corresponds to a share, i.e. another user finds your content useful and shares it with their community. On the other hand, a FAV, corresponds to a 'like' on Facebook or Instagram, i.e. a way for users to indicate that they like a particular tweet.

At this point, we would like to make a distinction between users who have many followers and those who have few followers. A person with many followers, consequently will also have more interaction than one with few followers, but this does not imply that their content is better, therefore, we will define engagement as the arithmetic mean of the interaction variables: number of retweets and number of favourites, normalised by the number of followers of the user. The number of retweets and favourites used is the accumulated sum of user retweets and favourites in the dataset in question, i.e. the retweets or favourites that a user has received on the topic under study. This value, therefore, offers a contrast to other formulas seen in the literature (Baum, 2019), where engagement modelling is done globally for all user-generated content without taking into account that this content may belong to more than one topic. We find that it is closer to reality to consider the engagement of a user for a certain topic, rather than the global engagement since a user can have experience in different areas. For example, a user can tweet about Artificial Intelligence (A.I.) with little success, and at the same time, about a sport that he practices, getting a lot of interaction in the tweets related to the sport. If we are analysing a topic related to A.I. and consider the global engagement of the user, we may have a bias that tells us that the user is relevant to A.I., when he or she is not.

Mathematically for each user $u \in U$, their engagement, denoted as ϵ, is calculated with the following formula:

$$\epsilon(u) = \frac{\frac{nFavsInTopic}{nFollowers} + \frac{nRtsInTopic}{nFollowers}}{2} \tag{1}$$

In this way, we achieve to increase the engagement of a user that has generated very useful content about the topic under analysis. For example, let us suppose the following simple example: *user1* with 20 Followers, 50 Favs, 30 Rts, and *user2*, with 1000 followers, 100 Favs and 120 Rts. If we apply the formula (1), we will have $\epsilon(user1) = 2$ and $\epsilon(user2) = 0.11$. So *user1*, will have more useful content than *user2*, because despite having less followers, they share more content in proportion to *user2*, who although having more followers, they do not interact as much. To pass the filter, we will select those accounts whose ϵ value is higher than the mean of all the engagement values.

It is necessary to mention that the modelling of engagement is very complicated, since the system can be susceptible to mark relevant users that interact a lot in the social network without caring about the content, although normally, the content that is relevant is shared. This is much more accentuated if we are in the professional field, where networks such as Twitter are often used to share and find research, results or studies. It is here where the cascade filter comes into play, because the previous expertise filter has already discarded non-professional accounts, so the system is less sensitive to this problem of sharing less valuable content.

Table 3
Machine specifications.

Component	Features
CPU	2 GHz Intel Core i5 with 4 cores
RAM	16 GB 3733 MHz LPDDR4X
VRAM	Intel Iris Plus Graphics 1536 MB
Hard disk	SATA SSD de 512 GB

3.4. Credibility filter module

The last filter is based on the credibility of the user. The user's credibility on a social network is intimately related to his or her popularity. That is, an account becomes popular because many other accounts believe it and therefore follow it and share its content. In other words, we can model credibility for our filter, based on an arithmetical mean of the Twitter values that are related to popularity. These values are: the number of followers, the number of public list in which the user appears, the number of retweets and the number of favourites. In the literature, other works closely related to the NOFACE framework use a standardised linear calculation of variables such as the number of followers, favourites, retweets and mentions. We have preferred to give importance to the lists, as opposed to the mentions, because the mentions are not necessary a good indicator as they can be mentions of anger or reproach, while the lists, have demonstrated in solutions like Cognos (Ghosh et al., 2012) offering good results. According to this, mathematically for each user $u \in U$, their credibility, denoted as ζ, is calculated with the next formula:

$$\zeta(u) = \frac{nFollowers + nLists + nRetweets + nFavs}{4} \tag{2}$$

To pass this last filter, the value must be above the mean of all ζ values. After applying this filter, the system will capture those user accounts related to the topic under study, whose content is usually interesting and who also have a wide popularity and credibility in social networks.

4. Framework evaluation

In this section we will go into detail in the experimentation carried out with the NOFACE framework as well as its application to a real problem. It is worth mentioning that all the code has been programmed in Python 3 and that the tests, the development and the application to a real problem have been carried out with the equipment whose specifications are shown in Table 3. The equipment is a non-professional laptop, which shows that the potential of certain techniques such as word embeddings can be democratised, and that a useful system does not necessary have to use large processing clusters to obtain a meaningful result.

4.1. Datasets

To check that the framework performs properly, it has been applied to two real problems, one relating to COVID-19 and the other to the US elections in November 2020. The datasets used for the experimentation have been released on Diaz-Garcia, Ruiz, and Martin-Bautista (2022). In this repository, the source code of the algorithms will also be released.

4.1.1. COVID-19 dataset

The disease caused by the new Coronavirus (Sars-Cov-2) (Zhou et al., 2020), first reported in Wuhan (Huang et al., 2020) in December 2019, now affects the entire world and is considered one of the largest pandemics in the history of mankind. The virus is present in all inhabited areas of the world, and has caused millions of infections and millions of deaths. Europe is currently one of the epicentres of the pandemic and is likewise one of the territories that is allocating the most resources to research into the new disease. One of the ways of research related to the pandemic, lies in the automatic processing of information related to the virus, because it is necessary to have systems that allow us to obtain truthful, summarised and useful information from those channels where the disease is reported and talked about.

One of these channels is Twitter, where there are millions of people talking about COVID-19, associated pathologies, virus mitigation measures, prevention measures or means of propagation. Being able to process this data correctly involves dealing with a lot of irrelevant information. Currently, the tweet dataset (Lamsal, 2020) related to COVID includes more than 700 million entries, which makes it a perfect candidate for testing our algorithm.

Our problem, uses a part of that dataset, specifically the tweets of the first week of the pandemic, which goes from March 20, 2020 01:37 AM to March 26, 2020 12:46 PM. The total number of tweets that have been taken into consideration for the experiment is 7 293 933 with 34 features for each of the tweets.

4.1.2. November 2020 US elections dataset

The elections on 3rd November 2020, pitted Democratic candidate Joe Biden against Republican, Donald Trump. During the days leading up to the election and up to election day, using Twitter's streaming API, we obtained a database of tweets. To filter the tweets related to the election, we saved those that used any of the following hashtags *#election2020*, *#november3*, *#2020election*, *#vote2020*, *#votebidenharris*, *#votedonaldtrump*, *#biden2020*, *#trump2020*, *#democrats*, *#republicans* or *#election*. For this use case, we have selected a part of the complete database. Specifically, the number of tweets taken into consideration for the experiments carried out in this paper is 2 118 180. These tweets are from 28 October at 01:55 PM to 30 October at 4:44 PM.

4.2. A use case in big tweet datasets

Lets assume that we need to apply techniques such as clustering, association rules or LDA to obtain valuable information about our dataset. We know that much of the content in social networks is noisy, data without value. We are also aware of the volatility and speed of data generation in social networks. We see that the number of tweets generated on a topic in a day exceeds the million. If we extrapolate this to a week or a month, the volume of data begins to be unmanageable and, in addition, the vast majority of these data will have no value for our analysis. This is where the NOFACE framework comes in, allowing the reduction of the data keeping only what really adds value to our analysis.

The objective of our use case is (among others) to apply data mining to obtain valuable information about COVID-19 or elections. It is about obtaining, for example, the most representative topics regarding virus containment measures, in the case of `COVID-19`, or clusters of tweets from independent, non-party biased sources of information in the case of `Elections`. The experimental design will deal with the comparison of results with the application of the NOFACE framework proposed in this paper and without its application. Therefore, the application of the framework in these use cases is intended:

- To reduce the amount of data thus favouring the computation time of the subsequent data mining algorithms.
- To help reduce the irrelevant information present on social media by maintaining only those instances with a high reputation and relation to the domain of the problem.
- To demonstrate that the content reduction of the algorithm improves the results in the subsequent data mining processes, in this case, clustering, association rules and LDA.
- To demonstrate that the NOFACE framework obtains topic relevant accounts.

4.2.1. Robustness checks

For the robustness checks of the system and to verify that the filters work properly, we have checked various factors inherent to the algorithm such as time, percentage of content reduction, and localised expert and reliable users. These concepts are linked to the proper functioning of the algorithm, since what the algorithm seeks to do is to reduce the amount of data in subsequent analyses. This explains why our approach obtains the same or better results in subsequent data mining processes.

For the evaluation of the results and improvements provided by NOFACE in conjunction with other data mining techniques, specific methods are used for each evaluation. Specifically the coherence in LDA (Röder, Both, & Hinneburg, 2015) and the silhouette coefficient in clustering (Rousseeuw, 1987). The silhouette coefficient in clustering measures how well defined the clusters are, while the coherence coefficient in LDA measures the relationship of terms within the same topic. We consider that these two robustness measures are the most suitable for our experimentation since they are the most widespread in the literature with respect to clustering and LDA. Regarding association rules we will rely on the number of obtained rules for a given confidence threshold, as well as the time to create transactions and obtain rules.

In addition, a graphical interpretation of the results is carried out using visualisation techniques. The interpretation of the visualisation and the results will be detailed in the next section.

4.2.2. Experimental results

The framework starts taking as input a set of words some related to health or to independent journalists, depending on the dataset used. The exact words are: *doctor, medical, researcher, medicine, epidemiologist* and *clinical* in `COVID-19` use case, and *communicator, nonprofit, truth, journalist* and *analyst* in the `Elections` use case. The mean score for passing the engagement filter has been set at 4 and 6, in the case of `Elections` and `COVID-19` respectively, whilst for passing the credibility filter the mean of the values has been 0.01 in both cases. These thresholds are defined by the mean engagement value obtained by the engagement analysis module and the mean credibility value obtained by the credibility analysis module of the NOFACE framework as explained in Sections 3.3 and 3.4. The results in terms of execution time, located users, final dataset size and percentage of content reduction can be seen in Table 4. Looking in detail Table 4, we have a comparison between the results when applying the pre-processing seen in Section 3.1 and the results when applying this pre-processing in conjunction with the NOFACE framework. The times are longer in the latter case, since we add one more layer of pre-processing, namely the NOFACE filter.

Looking at the content reduction, we see that in the cases where NOFACE is applied, the reduction is 99%. While in the cases in which the pre-processing is simply applied, we have a reduction of 77% which corresponds to the cleaning of retweets or tweets composed only of empty words, links or mentions.

Table 4
Results and intervals for each configuration and dataset.

Dataset	Configuration	Elapsed time	Located users	Original size	Final size	% reduction
COVID-19	pre-processing	[29 min 27 s–32 min 54 s]	1 047 496	7 293 933	1 652 975	77%
Elections	pre-processing	[5 min 49 s–6 min 41 s]	388 688	2 118 180	568 640	73%
COVID-19	pre-processing + NOFACE	[58 min 3 s–1 h 1 s]	[556–758]	7 293 933	[3 050–4 162]	**99%**
Elections	pre-processing + NOFACE	[17 min 33 s–23 min 32 s]	[995–1 026]	2 118 180	[4 108–4 241]	**99%**

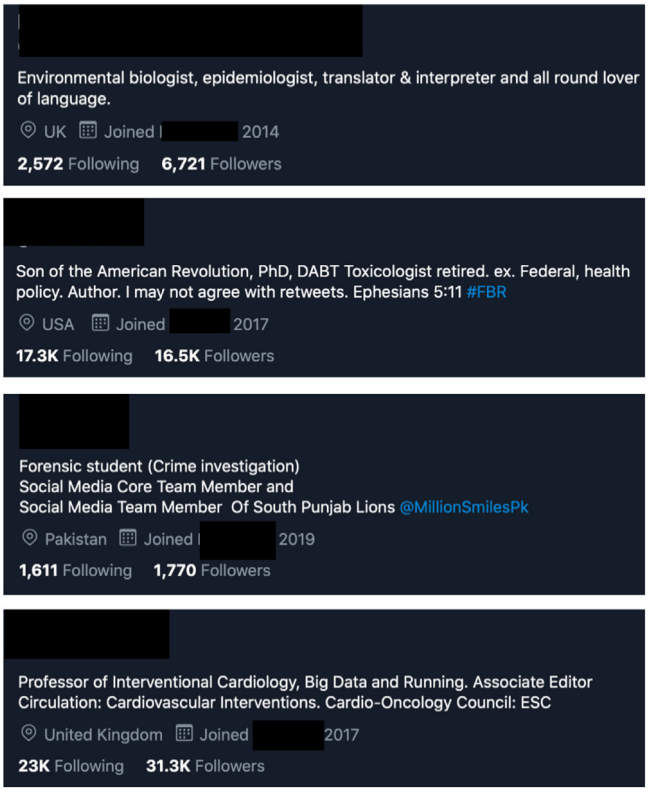

Fig. 2. Some anonymised profiles selected by NOFACE in the COVID-19 dataset.

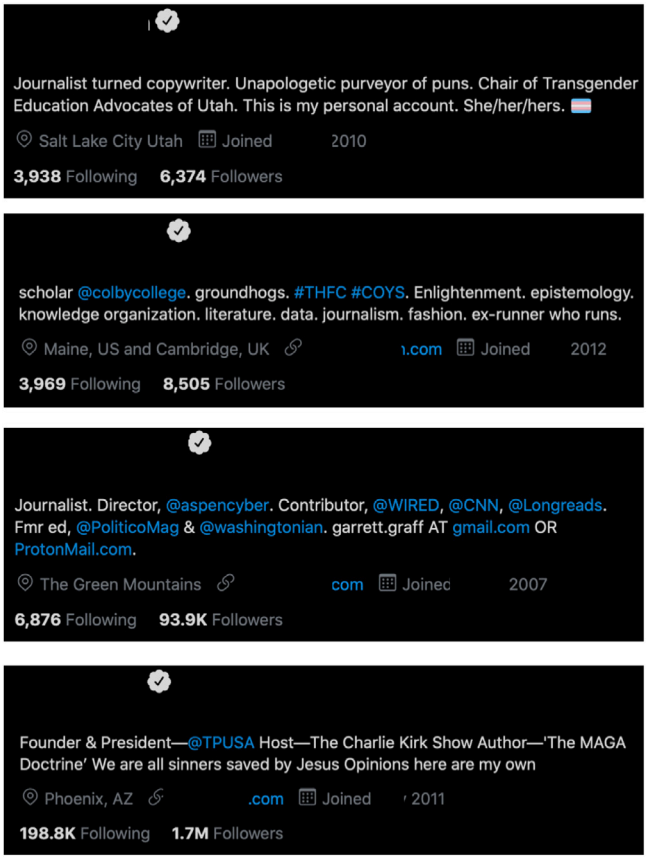

Fig. 3. Some anonymised profiles selected by NOFACE in the elections dataset.

We can conclude that the objective of content reduction is achieved and, moreover, the accounts that the algorithm considers relevant and interesting are actually correlated with the domain under study. Figs. 2 and 3 contain some of the accounts that have been considered as relevant by the algorithm, showing that they are profiles with a great reputation in relation to the subject matter of health or journalism. According to `COVID-19` use case, it is interesting to mention, that as shown in the picture, the profiles, except for the first one, contain words related to health not introduced in the first step of the process. The algorithm, using word embeddings, has learned how relevant they are to the topic and therefore takes them into consideration to enrich the search space. On the other hand, if we look at the use case of `Elections`, it is very interesting to see how a large number of accounts selected by the algorithm as truthful are accounts that Twitter has already verified, which is a great quality marker for the accounts filtered by the NOFACE framework.

Regarding execution times, the complete NOFACE framework always takes values ranging from 17 min to 1 h. On the other hand, the standard pre-processing applied to the control dataset takes 5 to 32 min to complete. Although it may seem that the time is slower with the NOFACE framework, we must considered that in the later stages that data mining algorithms could be applied (LDA, association rules and clustering in our case) we will have less amount of data to process, so the complete processing pipeline will take less time using the NOFACE framework. Table 5 shows the results in terms of time when the NOFACE framework is applied and not (where only pre-processing has been applied).

In the case of clustering, we can see very similar results to those seen in Table 4. In this case, when using NOFACE the time increase is in the order of milliseconds. Thus, the improvement is not very evident in the case of clustering, since the clustering algorithm used does not spend much time in the case of pre-processing either. Even so, in this case of using clustering, executions on the processed text took around 100 to 300 ms, while in the case of unprocessed text, we are dealing with the range of 5 to 10 s of execution. The reduction in proportion is considerable, much more if we extrapolate it to a problem with larger datasets. At this point, it is necessary to note that the clustering algorithm has a parameter which is the number of features it will use. In all experiments and configurations, this number of features is set to 50 000. Therefore, in both cases (with NOFACE and without NOFACE) only a part of the data is taken into account to perform the clustering computation.

In the case of obtaining topics with LDA, we can see an improvement in the total execution pipeline time. The ranges of pre-processing, applying NOFACE and LDA will always be lower than those of pre-processing and applying LDA directly. The reduction and cleaning of input data to the LDA algorithm performed by the NOFACE framework leads to a reduction of the total execution time of more than 30 min

Table 5
Elapsed time of the experiments with and without NOFACE.

Configuration\Dataset	COVID-19	Elections
Pre-processing	[29 min 27 s–32 min 54 s]	[5 min 49 s–6 min 41 s]
Pre-processing + Clustering	**[30 min 24–33 min 58 s]**	**[6 min 7 s–7 min 1 s]**
Pre-processing + LDA	[1 h 34 min 15 s–1 h 37 min 32 s]	[25 min 36 s–27 min 39 s]
Pre-processing + Apriori	[–]	[–]
Pre-processing + NOFACE	[58 min 3 s–1 h]	[17 min 33 s–23 min 32 s]
Pre-processing + NOFACE + Clustering	[58 min 4 s–1 h 1 s]	[17 min 32 s–23 min 33 s]
Pre-processing + NOFACE + LDA	[58 min 15 s–1 h 16 s]	[17 min 55 s–24 min 1 s]
Pre-processing + NOFACE + Apriori	[58 min 42 s–1 h 42 s]	[18 min 22 s–24 min 30 s]

in some cases. In Table 5 we can see that in both use cases, the best execution times for LDA are obtained when using the NOFACE filter.

One of the most interesting results can be obtained with regard to association rules mining. We have employed one of the most used algorithms for mining association rules, called Apriori. In this case, the algorithm cannot finish the execution due to the high amount of items and transactions to process. On the other hand, when filtering with the NOFACE framework, we obtain the rules in just a few seconds. This highlights the value of these filtering techniques for the use of subsequent algorithms in which the volume of data is a serious problem, like in the case of the Apriori algorithm.

Taking into consideration these results we can conclude that although a priori the NOFACE framework takes more time, if we considered the complete data processing pipeline, that is, in conjunction with other data mining techniques that could be interesting to apply for a complete analysis, it improves considerably the execution times.

4.2.3. Clustering results

Clustering, as far as texts are concerned, tries to find which documents are more similar to others, by placing them in the same cluster. In the Twitter domain, it tries to find out which tweets are more similar to others in terms of content, which has great implications in the process of summarising information, searching for influencers or categorising accounts, for example. Since it is one of the techniques widely used in text mining, we are going to apply K-means on the dataset filtered with NOFACE and the complete dataset. The characteristics that fed the clustering algorithm, correspond to a TF-IDF vectorisation of the document. To choose the number of clusters to search for, we have carried out an analysis using the sum of quadratic error (SSE). The value of clusters (k) used for experimentation, given by the SSE value, was 11 in the `Elections` use case and 13 in the `COVID-19` use case.

To graphically compare the obtained results we have represented them by means of a t-distributed stochastic neighbour embedding (TSNE) graph (Maaten & Hinton, 2008).

In Figs. 4 and 6 the results of applying the clustering algorithm without filtering are shown and in Figs. 5 and 7, the results in the case of filtering using NOFACE.

One of the first things we can observe in the TSNE graph is that in the case of the NOFACE results, we have more dispersion between the clusters. This is a clear symptom that good accounts have been selected in which the features are very differentiated. In the case of not applying NOFACE, we have more overlap between clusters, and the silhouette coefficient is of a worse degree.

Following with the analysis of the TSNE graph for the COVID-19 use case, in Fig. 4, the blue cluster is very dispersed over the whole area of the graph, while in Fig. 5 we can see that the majority cluster (in this case the purple one) is quite well defined, as also are the green, light green, red, magenta, dark blue, light blue, yellow, pink and orange ones. In this way, we can see how the application of NOFACE, has greatly improved the execution of the clustering algorithm, because in Fig. 4 it is complicated to identify more than 6 clusters (yellow, red, light blue, orange, red and magenta). This visual analysis is also corroborated by the calculation of the silhouette coefficient. Specifically, in the case of COVID-19, the silhouette coefficient is 0.0095

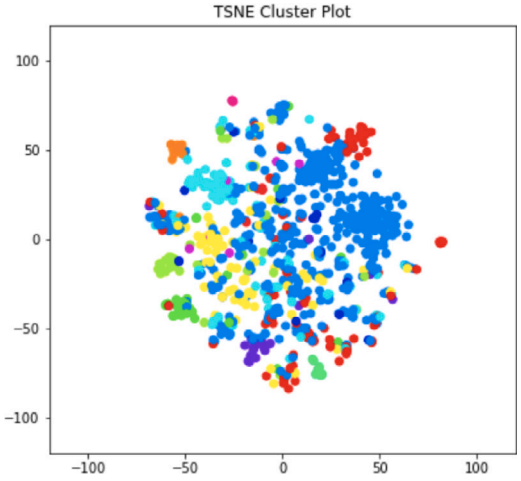

Fig. 4. COVID-19 use case: TSNE plot without NOFACE.

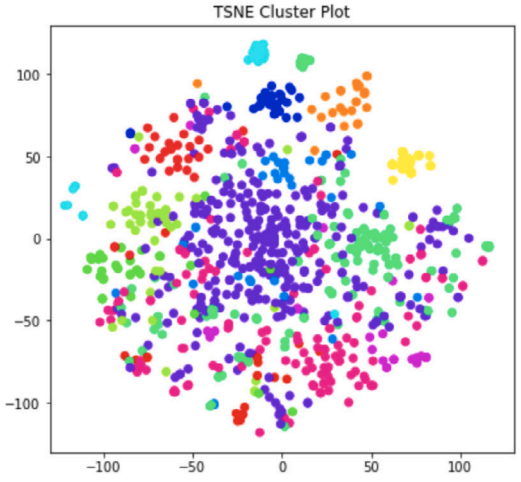

Fig. 5. COVID-19 use case: TSNE plot with NOFACE.

in the case of using the NOFACE framework and 0.0069 in the case of not using it. This coefficient gives us a value on how the clusters are differentiated from each other. This indicates that by applying the NOFACE framework, we have more differentiated and higher quality clusters. Although the improvement is not of a high degree, we can conclude that we have managed to reduce the dataset to one of a better quality with less data, so less computation time and easier interpretation improve the results.

In the case of `Elections` an improvement in the number of identifiable clusters can also be observed, although in this case this improvement is less evident than in the case of `COVID-19`. In this case, we can conclude that with such similar results, there is no significant loss of information when applying the NOFACE framework. It is also necessary to point out that being a political dataset and already

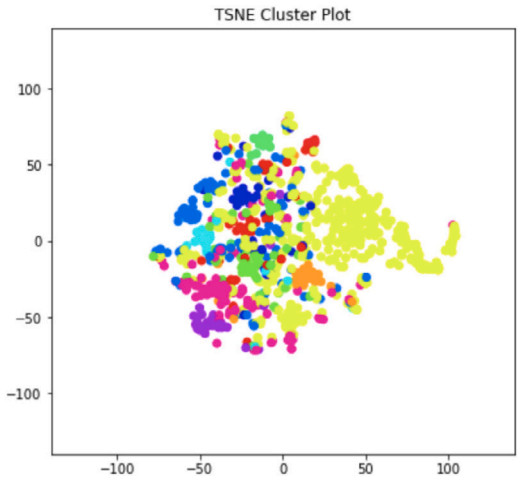

Fig. 6. Elections use case: TSNE plot without NOFACE.

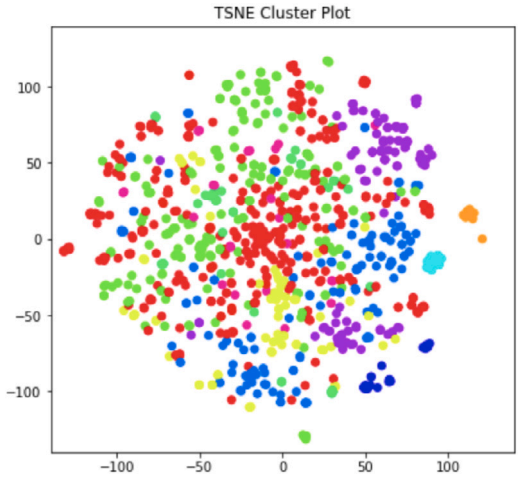

Fig. 7. Elections use case: TSNE plot with NOFACE.

filtered by hashtags related to the elections, the dataset will contain very polarised opinions based on ideology. This fits perfectly with the background of clustering, so as we can see in the results, in both cases the outcomes are generally strong. On the other hand, Fig. 7 shows how applying NOFACE makes the clusters more differentiated from each other. This analysis is reinforced by the silhouette coefficient, which is 0.12 when using NOFACE and 0.01 when not using it. Again, we see how the clustering algorithm offers better results if used in conjunction with the NOFACE filtering framework.

Another advantage that can be gained from using NOFACE is that by generating more cohesive clusters with less data, better content-based labelling of clusters can be carried out so that these can be used as classes to be applied to possible supervised classification problems.

4.2.4. LDA results

The LDA process seeks to obtain those topics that are being talked about in social networks. In our case, we tried to seek what are the main general topics about COVID-19 and elections in Twitter.

For example, a scenario for the application of the LDA on `COVID-19` would be to see if there is any kind of contradictory information, or relevant information for COVID-19 measures. In the case of the `Elections`, for example, it might be interesting to get related topics by state, to see what people are concerned in one state or another, or what is being discussed in the independent press. The process of obtaining topics takes the information from a bag of words

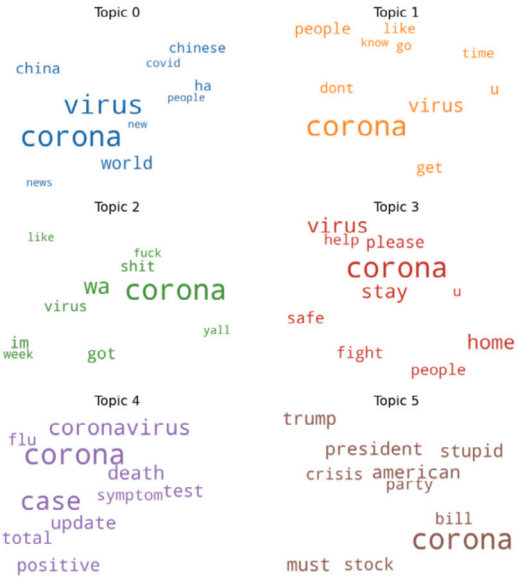

Fig. 8. Topics from the COVID-19 dataset processed without NOFACE.

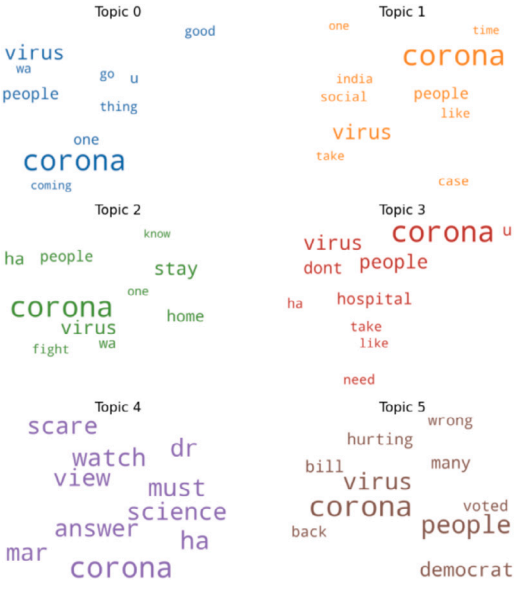

Fig. 9. Topics from the COVID-19 dataset processed with NOFACE.

generated from the text of the tweets. The number of topics was set to 6.

In Figs. 8 and 10 we can see the most representative words related to the 6 topics obtained by the LDA algorithm on the dataset that does not use NOFACE. On the other hand, Figs. 9 and 11 show the most representative topics obtained over the set of tweets filtered using NOFACE.

According to the figures, we can see how both outputs of the LDA algorithm contain very similar information, which shows that the NOFACE framework did not lose information and can therefore be very useful to keep those tweets and accounts that really add value. This analysis can be supported by coherence results, which give a value of 0.305 for COVID-19 using NOFACE, and 0.271 without using NOFACE. In the case of Elections, the improvement is less evident and the coherence results hardly fluctuate from one experiment to another. This leads us to reinforce the conclusion that there is no loss of important information, while there is a reduction of invaluable information and

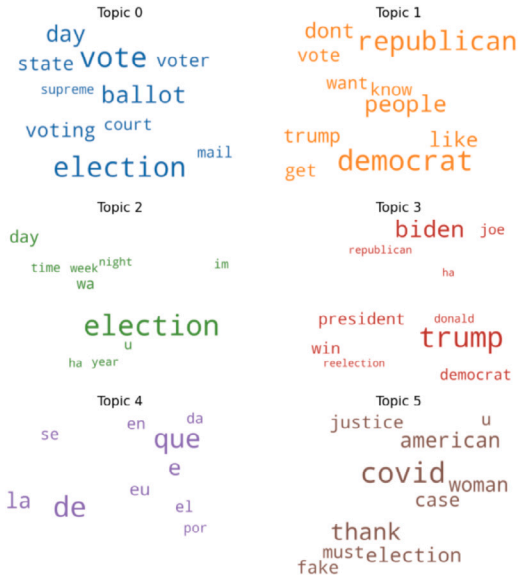

Fig. 10. Topics from the elections dataset processed without NOFACE.

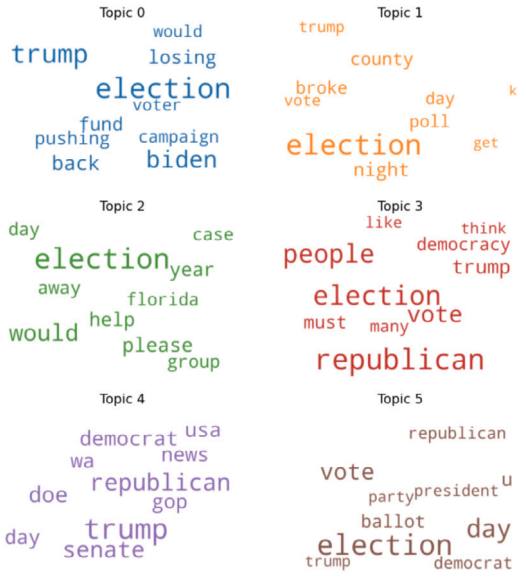

Fig. 11. Topics from the elections dataset processed with NOFACE.

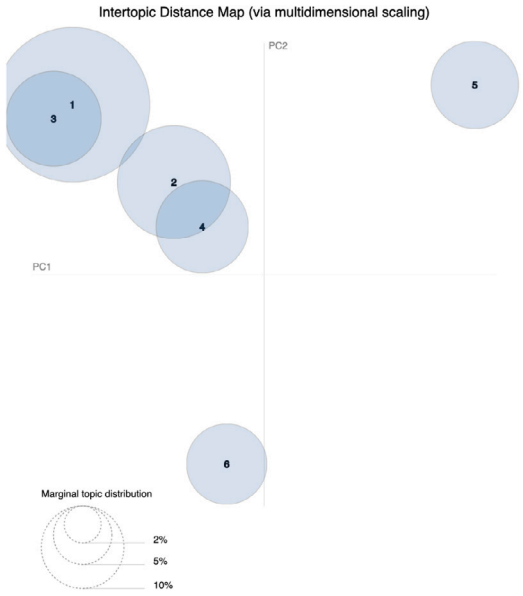

Fig. 12. Intertopic Distance Maps from the COVID-19 dataset processed without NOFACE.

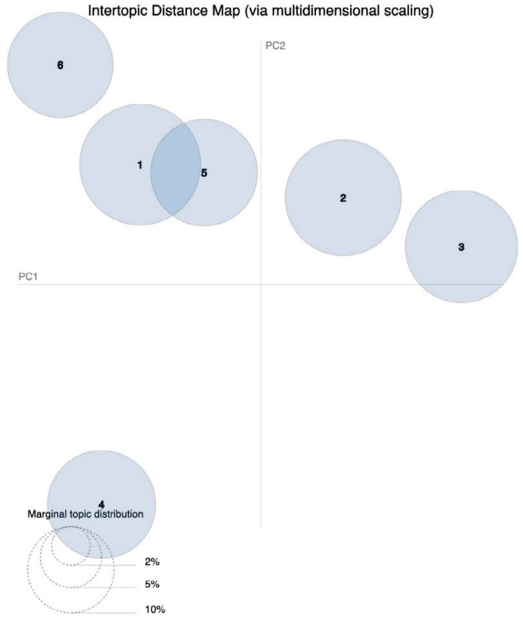

Fig. 13. Intertopic Distance Map from the COVID-19 dataset processed with NOFACE.

noise. Our analysis is focused on documents (tweets). Therefore, we have also experimented with the alpha hyper-parameter of the LDA algorithm, which can adjust sensitivity with respect to document-topic density or document-topic distribution. Specifically, we have used the symmetric and asymmetric values for each experiment. In the case of COVID, the best result in terms of coherence is obtained for the filtered dataset, with asymmetric alpha, obtaining a coherence value of 0.365. In the Elections dataset, the results are similar with both configurations.

In a more subjective analysis, we could even see words with more sense and relation with the COVID-19 or elections in Figs. 9 and 11 respectively. For example in the Election use case of topic 2, the text filtered by NOFACE (Fig. 11) contains related content about the state of Florida and the Democrat and Republican parties, which was a disputed and swing state until the last moment.

Finally, we have also added a display layer using graphics of topics according to Chuang, Manning, and Heer (2012). This graph is useful to show how the topics would be distributed in a 2D graph using principal components. Figs. 12 and 14 show the Intertopic Distance of results without applying NOFACE, whilst 13 and 15 are the results using the NOFACE framework. In the case of `COVID-19`, we see that there are clearly better results in the case of filtering using NOFACE, since we have more differentiated and dispersed topics, i.e. we have less intratopic overlap. Another analysis that can be distilled from the graphs is how the topics (circle size) are more homogeneous in the case of the NOFACE filtered datasets. This indicates that there is a better distribution of words between topics in this case, than if we compare it with the use cases without using NOFACE. On the other hand, in the `Elections` case we have a very similar overlapping and situation of the topics. Again we have a similar situation to the one we had in clustering. In this case we have obtained very similar results processing [4108–4241] accounts selected by NOFACE instead of 388 688 with traditional pre-processing. This brings a remarkable improvement in terms of performance, maintainability and analysis capability.

Fig. 14. Intertopic Distance Maps from the elections dataset processed without NOFACE.

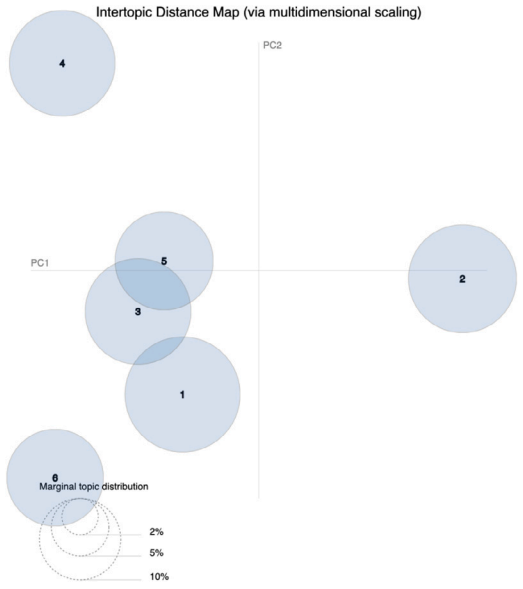

Fig. 15. Intertopic Distance Map from the elections dataset processed with NOFACE.

4.2.5. Apriori algorithm results

In the case of the association rules we cannot compare the results of both experiments because in the case of the unfiltered dataset the algorithm did not finish, as the explosion of frequent itemsets combinations is too high. In order to have some kind of comparison, we have carried out the experimentation by obtaining a random sample of the unfiltered dataset of the same size as the one resulting after filtering the dataset with NOFACE. Support values of 0.1 and 0.01 have been taken, for a confidence value of 0.6. For the support threshold value of 0.01 using the NOFACE filter, a total of 130 rules were obtained using the `COVID-19` dataset and 33 rules in the `Elections`. On the other hand, for the sample of the unfiltered dataset, this number is reduced to 105 in the case of `COVID-19` and 9 in the case of `Elections`. For the value of 0.1 for minimum support, only 1 rule was obtained in all cases except in the case of the random sample of the unfiltered `Elections` dataset where no rules were obtained. These results are interesting because they demonstrate how the filtering produces a more cohesive dataset, since the support value is a direct indicator of the co-occurrence of items in a dataset, and getting more rules in the case of applying NOFACE indicates that certain terms are more likely to appear together.

5. Discussion

After the experiment and analysis of the different use cases, it is necessary to review the results and check if the considered objectives set (Section 4.2) have been achieved.

One of the goals was to eliminate irrelevant content from the dataset. At this point, we can conclude that the NOFACE framework complies perfectly, perhaps even being too restrictive. As we have seen throughout the previous section, it reduces the content coming from Twitter to a great extent, in addition, it limits it to the scope of research that could be desired at a certain moment, health or journalism in our case. In other words, the framework, takes benefit of credibility, engagement and what is more important the user's experience in a domain, offering very representative content and profiles. The reduction in number of examples is very large, going from millions of tweets to just a few thousand. However, we must bear in mind that this reduction in terms of examples is appropriate for Big Data environments in social networks, where many of the content is noise or content generated by accounts with no value for the topic in question. Therefore, if we analyse the result from the point of view of users with experience for the topic, and in a period of 2 days, the algorithm detects from 500 to 1000 relevant users among the available 10 000 users. Thus, analysing the content of those 1000 users with experience in the topic will be more efficient, than analysing 10 0000 who have simply given their opinion on the topic.

Another objective was to compare the execution times and to check if the framework introduces any computational improvements. In this case, there are conflicting results, because in the case of LDA and association rules, the framework obviously improves the execution times. In the case of clustering, the extra time involved in running the NOFACE framework makes it behave worse in terms of time. In this case, it is necessary to mention that both the standard pre-processing experiment and the NOFACE experiment employed 50 000 characteristics for the TF-IDF vector. Therefore this result is biased by this value, that at this point is a dimensionality reduction based on the characteristics of the TF-IDF. This means that at this point for both experiments (with and without NOFACE), the feature dataset used is considerably reduced, since the TF-IDF discards many features (words). Thus, we are not using the full dataset in the case of the experiment without NOFACE, just a selection of the best textual features guided by the TF-IDF. This makes the results more similar in terms of elapsed time. Even so, the clustering algorithm takes about 0,001 s to finish with the dataset processed with NOFACE, and from 19 to 64 s with the complete dataset, so the improvement achieved by the proposed filtering method is still evident.

The last goal was to demonstrate that the results of other data mining techniques on a dataset processed with NOFACE, would improve against a dataset that had not been processed with our framework. Throughout this section, we have seen how in the worst case, the result is very similar, which indicates that the framework really selects good profiles whose information is relevant, that is, there is no loss of relevant information. In other cases, we have seen how the framework improves considerably, as in the case of the clustering algorithm on the `COVID-19` dataset, where cohesive and differentiated clusters with fewer outliers were found. Finally, in the case of association rules, we have seen how the filter can help certain algorithms, sensitive to the amount of data, to function normally, improving the results in terms of obtained rules.

5.1. Challenges

One of the most important challenges of the NOFACE framework, as well as other similar systems, involves dealing with lies and noise in social media. We must bear in mind that nothing prevents a person from describing themselves in their biography as a doctor, researcher or engineer without actually being one. This makes the system very sensitive to these issues, and it is therefore a challenge to design an automatic system that is capable of discerning between real people and those who are not. The system proposed in this paper, as well as other proposals presented in the literature, offers more layers of analysis (engagement and credibility) trying to mitigate this problem, having as a premise that probably real Twitter followers will not share content from people of dubious biography or belonging.

Another of the great challenges to be faced by content-based systems and not by graph-based systems, as proposed in this paper, is that they can detect a fake influencer as relevant. The fake influencers are behind accounts in social networks with great statistics of interaction and content generation. This can lead a brand to think that it can be a great investment to hire this account to spread their products or services, but in reality it would be a waste of money because these accounts really generate interaction with accounts managed by bots and other non-real accounts, so there is no real interaction. Detecting these accounts requires network analysis, and according to Tsapatsoulis, Anastasopoulou, and Ntalianis (2019), they are usually egocentric accounts easily identified by network analysis algorithms based on centrality.

5.2. Contributions to literature

The main contribution of this paper to the literature has been the creation of a framework for the selection of relevant content and users in social networks. Throughout the paper it has become clear that social networks play an irreplaceable role in our daily lives, and in particular in many business processes. The review (Kumar et al., 2021) shows how users use social networks to inform themselves about products and services of various kinds. It also mentions several studies on how companies of different sizes use social networks to obtain information from users in numerous aspects (Chatterjee & Kar, 2020). It is in these points, where our framework takes special relevance and can help companies or individuals to filter the content of social networks to favour their subsequent stages of data analysis. With this filtering, the algorithm also achieves a reduction of the dataset to be processed.

Another contribution to the state of the art relates to misinformation. By selecting credible users, with experience in the sector and with a certain impact, we are also ruling out the misinformation component to a certain extent. Therefore the proposed framework helps to eliminate the misinformation present in social networks. In this sense, we are carrying out an elimination through user features (favourites, retweets) and information obtained from natural language processing of their biographies. Currently, there are other models specially designed for these tasks that use classification algorithms or deep learning models (Mahir, Akhter, Huq, et al., 2019) such as recurrent neural network models and LSTM, to classify whether something is misinformation or not. As we have seen throughout the paper, these models need prior training, something that makes them sensitive to changes. So we find that there is a need for systems, such as the one proposed throughout this paper or others in the literature, based on content features (Wu, Liu, Liu, Wang, & Tan, 2016) and filters that allow to narrow down the amount of false content in a way that does not require large databases and training.

Finally, three comprehensive use cases of data mining in conjunction with the NOFACE framework have been provided to the literature. In the case of the paper, clustering, association rules and LDA have been used. It has been demonstrated how the framework improves clustering results in terms of silhouette and cohesion of the clusters. As for LDA, the topics obtained are of better quality in terms of coherence. The literature (Joung & Kim, 2021) highlights the need for these pre-processing techniques to improve the performance of algorithms such as topical detection algorithms (LDA). Also, in this paper it has been highlighted how the use of efficient pre-processing can help to improve the execution times of a complete data mining pipeline. It has been shown that the filter can be of special interest in those algorithms where the number of data can make them fail or the execution time is very inefficient, as in the case of the Apriori algorithm (Al-Maolegi & Arkok, 2014).

6. Conclusion and future work

The present work has proposed a new framework for filtering irrelevant content on social media, demonstrating its usefulness as a technique for pre-processing data before applying other data mining techniques such as clustering, association rules or LDA. Two of the most widespread robustness metrics in these techniques (silhouette and coherence) have been used on the datasets filtered by the NOFACE framework. Based on these metrics, it has been shown that the framework does not lose information and improves the results obtained. Additionally the proposed framework can be used with a wide range of data mining algorithms, being specially appropriate on those that may be limited by data size and those that may be sensitive to noise for a given type of analysis.

During the development and research, a study of the state of the art in the subject has been carried out. The use of advanced text mining techniques based on word embeddings has also been highlighted. This is, as far as we know, the first contribution that applies these techniques to compute expertise on a topic.

Also, the potential of the framework has been highlighted on two real problems of tweets relating to COVID-19 and the 2020 US elections, on which the consequent reduction of number of examples without loss of information has been demonstrated, even improving the results obtained using the complete dataset. In short, the paper:

1. Offers a new framework for irrelevant content reduction in social networks based on iterative filters. It also helps to reduce misinformation as it is usually issued by inexperienced or low credibility users in a particular sector, being these discarded by NOFACE. Iterative filters address the problem in a very strict way and can alleviate the problem of lies about professional experience on social networks.
2. It introduces an algorithm for locating experts in social networks through the use of word embedding. It has been demonstrated that the algorithm is feasible and can be used as a pre-processing step prior to other data mining applications.
3. It proposes an interpretable and easily understandable solution to the problem of detecting user-generated content useful for a given topic or analysis.
4. Provides two detailed interpreted use cases to support the use of the framework in conjunction with other data mining tasks. In these use cases it has been demonstrated that there is no loss of information and improved results in terms of computation time and robustness.

The proposed framework opens up future channels of development that are closely linked to the challenges seen in Section 5, like the study of how sensitive is the system towards lies and egocentric networks generated by fake influencers or false credibility, so being able to identify these issues would considerably improve the system. It is also necessary to mention the opposite case to the one described above, since maybe the system is not considering users who are very influential in their field, but who have hardly any presence in social networks. That is, the framework in its current state is very restrictive, so being able to locate the low statistical but really good accounts would be a

great improvement and a future path of development and research as well.

Although two unsupervised use cases have been provided in the use cases, the framework could also be used in supervised methods. A possible future application in this sense would be to filter a large dataset of topic-related data into useful or truthful information and non-relevant or fake information using the NOFACE framework. Then, using these resulting labelled datasets to train a classifier, for example based on deep learning, that allows us to determine whether a new tweet is truthful or not.

Finally, there is the possibility of extending the system to a purely streaming environment, where related words could be mutated by time windows and a list of expert users would be maintained over time, who could cease to be experts if their engagement or credibility levels drop.

CRediT authorship contribution statement

J. Angel Diaz-Garcia: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft. **M. Dolores Ruiz:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Maria J. Martin-Bautista:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have link the data to the manuscript.

Acknowledgements

Funding for open access charge: Universidad de Granada/CBUA. The research reported in this paper was partially supported by the COPKIT project under the European Union's Horizon 2020 research and innovation program (grant agreement No 786687), the Andalusian government and the FEDER operative program under the project BigDataMed (P18-RT-2947 and B-TIC-145-UGR18). The paper is part of the NOFACEPS project (PPJIB2021-04) of the University of Granada's internal plan. Finally the project is also partially supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150). All authors approved the final version of the manuscript.

References

Abu-Salih, B., Wongthongtham, P., Chan, K. Y., & Zhu, D. (2019). CredSaT: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor. *Journal of Information Science*, *45*(2), 259–280.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB, Vol. 1215* (pp. 487–499). Citeseer.

Al-Maolegi, M., & Arkok, B. (2014). An improved apriori algorithm for association rules. arXiv preprint arXiv:1403.3948.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Alrubaian, M., AL-Qurishi, M., Alrakhami, M., Hassan, M., & Alamri, A. (2016). Reputation-based credibility analysis of Twitter social network users: Reputation-based credibility analysis of Twitter social network users. *Concurrency Computations: Practice and Experience*, *29*, http://dx.doi.org/10.1002/cpe.3873.

Alrubaian, M., Al-Qurishi, M., Hassan, M. M., & Alamri, A. (2018). A credibility analysis system for assessing information on Twitter. *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 661–674. http://dx.doi.org/10.1109/TDSC.2016.2602338.

Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: managing misinformation in social media—insights for policymakers from Twitter analytics. *Journal of Data and Information Quality (JDIQ)*, *12*(1), 1–18.

Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, *56*(4), 1234–1244.

Batra, J., Jain, R., Tikkiwal, V. A., & Chakraborty, A. (2021). A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. *International Journal of Information Management Data Insights*, *1*(1), Article 100006.

Baum, A. (2019). Scraping Twitter user data using google and tweepy. https://towardsdatascience.com/use-google-and-tweepy-to-build-a-dataset-of-twitter-users-cbfd556493a9 [Online; accessed 18-January-2020].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Canini, K. R., Suh, B., & Pirolli, P. L. (2011). Finding credible information sources in social networks based on content and social structure. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 1–8). IEEE.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675–684).

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28.

Chatterjee, S., & Kar, A. K. (2020). Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from India. *International Journal of Information Management*, *53*, Article 102103.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74–77).

Cordeiro, P. R. D., Pinheiro, V., Moreira, R., Carvalho, C., & Freire, L. (2019). What is real or fake?-machine learning approaches for rumor verification using stance classification. In *IEEE/WIC/ACM international conference on web intelligence* (pp. 429–432).

Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. arXiv:1605.07891.

Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2022). NOFACEPS source code and data repository. URL https://github.com/ugritlab/NOFACEPS.

Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., & Gummadi, K. (2012). Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 575–590).

Hassan, D. (2018). A text mining approach for evaluating event credibility on twitter. In *2018 IEEE 27th international conference on enabling technologies: Infrastructure for collaborative enterprises (WETICE)* (pp. 171–174). IEEE.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506.

Joung, J., & Kim, H. M. (2021). Automated keyword filtering in latent Dirichlet allocation for identifying product attributes from online reviews. *Journal of Mechanical Design*, *143*(8).

Kaliyar, R. K. (2018). Fake news detection using a deep neural network. In *2018 4th international conference on computing communication and automation (ICCCA)* (pp. 1–7). IEEE.

Kang, B., O'Donovan, J., & Höllerer, T. (2012). Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on intelligent user interfaces* (pp. 179–188).

Kar, A. K., & Aswani, R. (2021). How to differentiate propagators of information and misinformation–Insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences*, *42*(6), 1307–1335.

Khoo, L. M. S., Chieu, H. L., Qian, Z., & Jiang, J. (2020). Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 34* (pp. 8783–8790).

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*(1), Article 100008.

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2021). Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management*, *58*(5), Article 102631.

Kuzi, S., Shtok, A., & Kurland, O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 1929–1932).

Lamsal, R. (2020). Coronavirus (COVID-19) tweets dataset. http://dx.doi.org/10.21227/781w-ef42.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 302–308).

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings. In *Twenty-ninth AAAI conference on artificial intelligence*. Citeseer.

Liu, Y., & Wu, Y.-F. B. (2020). Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, *38*(3), 1–33.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., et al. (2016). Detecting rumors from microblogs with recurrent neural networks.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1* (pp. 281–297) Oakland, CA, USA.

Mahir, E. M., Akhter, S., Huq, M. R., et al. (2019). Detecting fake news using machine learning and deep learning algorithms. In *2019 7th international conference on smart computing & communications (ICSCC)* (pp. 1–5). IEEE.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Molina-Solana, M., Amador Diaz Lopez, J., & Gomez, J. (2018). Deep learning for fake news classification. In *I workshop in deep learning, 2018 conference Spanish association of artificial intelligence* (pp. 1197–1201).

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv:1902.06673.

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-rnn based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), Article 100007.

Oehmichen, A., Hua, K., Amador Dííaz López, J., Molina-Solana, M., Gómez-Romero, J., & Guo, Y. (2019). Not all Lies are equal. A study into the engineering of political misinformation in the 2016 US presidential election. *IEEE Access*, *7*, 126305–126314. http://dx.doi.org/10.1109/ACCESS.2019.2938389.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review*, *34*(2), 133–143.

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications*, *540*, Article 123174.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). URL: http://www.aclweb.org/anthology/D14-1162.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on web search and data mining* (pp. 399–408).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Roy, D., Paul, D., Mitra, M., & Garain, U. (2016). Using word embeddings for automatic query expansion. arXiv:1606.07608.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*(2), 907–948.

Tsapatsoulis, N., Anastasopoulou, V., & Ntalianis, K. (2019). The central community of Twitter ego-networks as a means for fake influencer detection. In *2019 IEEE intl conf on dependable, autonomic and secure computing, intl conf on pervasive intelligence and computing, intl conf on cloud and big data computing, intl conf on cyber science and technology congress (DASC/PiCom/CBDCom/CyberSciTech)* (pp. 177–184). IEEE.

Wu, S., Liu, Q., Liu, Y., Wang, L., & Tan, T. (2016). Information credibility evaluation on social media. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 30*.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, *21*(2), 80–90.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273.