International Journal of Computational Intelligence Systems (2024) 17:93
https://doi.org/10.1007/s44196-024-00483-y

RESEARCH ARTICLE

A Flexible Big Data System for Credibility-Based Filtering of Social Media Information According to Expertise

Jose A. Diaz-Garcia[1] · Karel Gutiérrez-Batista[1] · Carlos Fernandez-Basso[1,2] · M. Dolores Ruiz[1] · Maria J. Martin-Bautista[1]

Received: 2 December 2022 / Accepted: 23 March 2024

Abstract

Nowadays, social networks have taken on an irreplaceable role as sources of information. Millions of people use them daily to find out about the issues of the moment. This success has meant that the amount of content present in social networks is unmanageable and, in many cases, fake or non-credible. Therefore, a correct pre-processing of the data is necessary if we want to obtain knowledge and value from these data sets. In this paper, we propose a new data pre-processing technique based on Big Data that seeks to solve two of the key concepts of the Big Data paradigm, data validity and credibility of the data and volume. The system is a Spark-based filter that allows us to flexibly select credible users related to a given topic under analysis, reducing the volume of data and keeping only valid data for the problem under study. The proposed system uses the power of word embeddings in conjunction with other text mining and natural language processing techniques. The system has been validated using three real-world use cases.

Keywords Social media mining · Pre-processing · Big data · Expertise · Credibility

Abbreviations

COVID	Coronavirus disease
NLP	Natural language processing
KNN	*K* nearest neighbors
TF	Term frequency
TF-IDF	Term frequency-inverse document frequency
NOFACE	NOise filtering according credibility and expertise
SRTD	Scalable and robust truth discovery
FAV	Favorite
RT	ReTweet
URL	Uniform resource locator
CSV	Comma-separated values
MR	MapReduce
RDD	Resilient distributed datasets
API	Application programming interface
AWS	Amazon web services

✉ Jose A. Diaz-Garcia
jagarcia@decsai.ugr.es

Karel Gutiérrez-Batista
mbautis@decsai.ugr.es

Carlos Fernandez-Basso
mdruiz@decsai.ugr.es

M. Dolores Ruiz
karel@decsai.ugr.es

Maria J. Martin-Bautista
cjferba@decsai.ugr.es

[1] Department of Computer Science and A.I, University of Granada, C. Periodista Daniel Saucedo Aranda, s/n, 18014 Granada, Granada, Spain

[2] Causal Cognition Lab, Division of Psychology and Language Sciences, University College London, London, UK

1 Introduction

Social media completely influence today's world [1]. They are used daily by millions of people to share information or obtain information about current topics. This use generates an immeasurable amount of data, which, when correctly analyzed, can be of great value to companies, organizations, or even the users of social networks themselves [2]. In the social media context, most of the data generated that may be of interest comes in the form of unstructured text, such as opinions about a product, reviews about a restaurant, or simply conver-

Published online: 15 April 2024

sation threads on a given topic. Due to its nature, this type of data has an associated problem since, being user-generated content, we can find incomplete data, false or irrelevant content, colloquial uses of words, syntactic errors, and use of emoticons or simplifications of words [3]. In addition to those problems, social media data usually have issues related to the volume of the dataset. Millions of tweets, posts, news, or images are posted to social networks daily. Traditional data mining techniques cannot process properly that volume of data or have serious problems doing it [4].

Therefore, to derive value from this user-generated content, we must apply efficient data pre-processing techniques that allow us to have quality datasets from which to derive value and knowledge. In this paper, we offer a new technique that addresses two issues related to processing user-generated textual content in social networks: data validity and data volume reduction. The proposed technique is essentially a pre-processing technique designed to obtain a reduced and valid set of data for a given topic from a big dataset of data from social networks. On this dataset, other data mining techniques can be applied, with special value for those in which the volume of data can derive efficiency issues.

The premise of our paper focuses is the fact that many professionals use social networks such as Twitter to disseminate their knowledge and opinions about their field of expertise. The data from these professionals will be those that contribute the most value to the desired analysis or use cases, and therefore, they are the ones that should be located. Given those biographies on social networks allow us to provide information about our work, likes, or interests, our system will focus on using this textual element to discern who might be interesting to take into account as an expert and who might not. To achieve this, the core of the system uses word embeddings and their potential to obtain semantic relations between words. For example, two words of similar meaning, such as *doctor* and *surgeon*, will have a very small distance in the low-dimensional vector space so that by distance subtraction, we can obtain those words that are similar to a given word. In this way, by comparing the words with the topic under analysis, we will have a broad relationship to the topic and, therefore, can be used to expand the search query and perform flexible filtering by topic of analysis. This paper is the final result of the work [5] where a first approach to the expertise filter was conducted and in which a comparative battery of experiments was generated to discern the underlying word embedding of the final algorithm. The research presented in this paper introduces significant advances compared to our previous paper [5], particularly with respect to the final algorithm presentation, validation, and system testing across diverse use cases. This paper showcases the refined and enhanced version of the algorithm, which effectively addresses the limitations identified in prior research. A key improvement lies in the selection of words using a threshold-based approach rather than relying solely on retrieving the top five most similar words. The algorithm demonstrates improved performance and delivers more accurate results by incorporating a threshold. This enhancement significantly enhances the algorithm's ability to capture relevant semantic-contextual representations of texts. Moreover, the most notable contribution of this research lies in its successful extension to a flexible version under the big data paradigm. Recognizing the ever-increasing scale and complexity of modern datasets, the algorithm has been optimized and tailored to process and analyze large volumes of textual data efficiently. The validation and testing of the system have been conducted rigorously across a variety of use cases, ensuring its robustness and generalizability. The experimental results highlight the algorithm's superior performance and its ability to provide accurate and explainable results, even when dealing with extensive and diverse datasets. The major contributions of our paper to the state-of-the-art are:

- The proposal of a flexible system capable of generalizing and adapting to a user's information needs at any given time. The system is based on filtering content based on experience, so it obtains content issued by people with knowledge in a given field.
- The development of the complete system under the Big Data paradigm using Spark [6, 7], highlighting its great usefulness in problems involving large data sets from social networks.
- The proposal of a pre-processing system that can be used to solve two of the most plaguing problems of Big Data: the validity and the volume of data.

To validate the system, a series of use cases and experiments have been proposed on a real dataset from Twitter. The filtering system proposed in this paper has been applied to different information needs, one related to COVID and health, one to politics, and one to sports. In all cases, it is shown how the system adapts flexibly to the user needs, obtaining valid data for that topic according to the users' experience in it, considerably reducing the volume of data and achieving a data sample that represents the topic with higher credibility, augmenting thus their quality and validity. For each use case, we demonstrate that our algorithm works properly using Named Entity Recognition (NER) [8, 9] and visualization based on tag clouds. In the visualization of the last layer of the process, we conduct an analysis that corroborates how the system can obtain valuable information starting from a non-quality dataset in an unsupervised way. It is crucial to emphasize that the credibility of our paper is intricately tied to the expertise and domain knowledge of individuals with professional backgrounds directly related to the topic at hand. In Section 3, we delve deeper into this notion, providing a

comprehensive explanation of the significance of expertise and its influence on the credibility of our research.

The rest of the paper is organized as follows: the next section focuses on the study of related work. In Sect. 3, we explain the algorithm proposed for experience-based filtering. In Sect. 4, we explain the Big Data architecture for the proposed system. In Sect. 5, we explain the experimental process carried out. Finally, in Sect. 6, we illustrate and validate the system with three real use cases. Final remarks and possible extensions are discussed in Sect. 7.

2 Related Work

As mentioned above, in this research, we propose a new data pre-processing technique based on Big Data and the power of word embeddings to solve two of the Vs regarding the Big Data paradigm: validity and volume. Many studies have addressed the problem of filtering information from social networks, most based on credibility-based dimensionality reduction techniques, either at content level [10–12], or user level [13–16] (see Table 1). Studies can be found that tackle the problem of filtering information through other approaches, such as fake news detection [17–19].

Considering that one of the main goals of the proposal is to develop a system under the distributed paradigm to be able to deal with large datasets from social networks, the literature on this topic has been reviewed.

2.1 Credibility-Based Information Filtering

In [10], the authors propose a supervised method to analyze the credibility of a given set of tweets from Twitter in an automatic way. The authors extract relevant features from tweets considered as trending topics to create the model. The features are extracted from the tweets, users' behavior, and citations to external sources. Finally, using all the features mentioned above, the model can classify the tweets as credible or not. It should be noted that, unlike our proposal, the authors use users' information, like biography, in a straightforward way.

Hassan [12] develops a text mining-based approach to assess the credibility of social network events automatically. The author uses a set of popular Twitter events manually annotated with different credibility ratings to build a model based on the Decision Tree classifier.

Canini et al. [11] address the problem of filtering credible information in social networks based on its content and structure. In this research, the authors detect (through experimental results) different factors affecting explicit and implicit credibility judgments in online social networks. Based on these results, they propose an approach to automatically identify and rank social network users according to their relevance and expertise for a given topic.

As stated, some approaches tackle the challenge of filtering credible information through fake news detection or similar tasks. That is the case in [19], where the authors propose a method for spam detection in emails based on the KNN classifier jointly with bio-inspired optimization techniques. In [18], a two-step-based method for fake news detection in social media is proposed. The first step comprises several pre-processing techniques in order to transform unstructured into structured data. The news in the transformed data is represented using the old-fancy Term Frequency (TF) feature representation. Finally, 23 classification algorithms are used to classify the news into fake or real.

Most of the approaches in this direction are based on supervised machine-learning techniques. Such is the case in [17], where the authors apply classic machine learning and deep learning algorithms for stance classification and rumor verification tasks. The obtained results conclude that classic algorithms outperform deep learning models for both tasks, and the information on stance does not enhance the rumor verification task.

Some works address the problem of filtering irrelevant content considering user-level information. In [13], Alrubaian et al. present a credibility analysis system for assessing information credibility on Twitter in order to avoid the increase of fake or malicious information. The proposal is based on four components that allow analyzing and assessing the credibility of Twitter tweets and users (considering their reputation and experience). The authors use four machine learning algorithms to showcase the feasibility of the system achieving a significant balance between recall and precision.

The same authors propose in [14] a novel approach that analyzes the user's reputation in the social network for a given topic. The proposal allows measuring the user's sentiment in order to recognize suitable and credible sources of information. The performance of the proposed reputation metric is evaluated with two machine learning algorithms, concluding that the approach can identify credible Twitter users.

The main distinction between our proposal and the previous approaches is how information is filtered. Existing approaches predominantly rely on user-generated content, such as posts, comments, or tweets, to assess credibility and relevance. While these approaches provide valuable insights into users' opinions and perspectives, they may not directly reflect users' professional expertise or work experience. In contrast, our proposal takes a different approach by leveraging users' biographies, which are explicitly intended to provide information about their work and professional backgrounds. By focusing on biographies, we can tap into a more direct and explicit source of information that is closely tied to users' expertise and qualifications. User biographies often include details about an individual's occupation, job title,

Table 1 Summary of related works

Methods	Type	Summary
[10]	Content level	Features are extracted from the tweet, user behavior and citations to external sources. Finally, using all the mentioned features, the model using the supervised method can classify the tweets as credible or not
[11, 17]	Content level	The authors of this study identify various factors that influence explicit and implicit credibility assessments in online social networks using experimental findings
[12]	Content level	To train a model based on a supervised method (Decision Tree classifier), the author utilizes a collection of popular Twitter events manually annotated with varying credibility ratings
[13, 14]		The goal of this proposal is to calibrate user sentiment and discern credible sources of information
[15]	User level	This proposal, which is capable of processing large data sets, considers the content and time factor to build a ranking of competent and meaningful users in the social network
[16]	User level	The approach is based on the fact that if a user is credible, their content is credible. The authors use word embeddings and Big Data methods to capture the semantics of the user's profile texts
[18, 19]	User level	These proposals use pre-processing comprising various techniques to transform unstructured data into structured data. Finally, classification algorithms are used to classify the news as fake or real

industry, and specialized skills. This information offers a deeper understanding of users' professional backgrounds and allows for a more targeted assessment of credibility in specific domains or topics.

Finally, it is worth noting a new trend closely related to our research, which focuses on filtering documents that have a high probability of being useless, non-credible, or containing misinformation. This trend aligns with our approach as we aim to filter the corpus of documents based on the credibility of the user who tweeted or created the content. However, this new trend focuses on filtering the content based on the source of information retrieval. In [20], the authors propose Vera, a system based on the sequence-to-sequence model T5 [21], which scores the usefulness of health documents for specific queries. Building upon this research, Zhang et al. [22] extend Vera to assess the usefulness of websites for health topics, discarding websites with misinformation during the retrieval process.

Similar to the aforementioned studies, our system has been developed to generalize and score the usefulness of documents for specific topics. However, our approach goes beyond the trained topic, as our system is flexible and can be extended to other topics without the need for retraining. This flexibility is achieved by leveraging and utilizing user information from the social network, enabling us to consider the credibility and expertise of users in assessing the usefulness and relevance of documents.

2.2 Big Data and Credibility-Based Information Filtering

Credibility-based information filtering has also been addressed from a Big Data perspective. Big data-based approaches include those works that deal with a massive amount of data and aim to reduce the amount of misinformation. Although the problem of filtering misinformation from large amounts of data has not yet received the attention it deserves [23], we can find some research in this direction.

That is the case in [15], where a Big Data-based solution is proposed. The system, called CredSaT (Credibility incorporating Semantic analysis and Temporal factor), considers the content and the temporal factor in order to build a ranking of proficient and significant 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.

Diaz et al. [16] present a user-centered framework for filtering irrelevant content in social networks to facilitate data mining techniques post-usage. The proposed system, called NOFACE (NOise Filtering According to Credibility and Expertise), also helps to reduce misinformation in social networks since it identifies credible and renowned members. The proposal relies on the fact that if a user is credible, its content is also credible. The authors use word embeddings to capture the semantics of the texts in the user's profile. Then, these word embeddings enable the creation of a group of relevant users based on their expertise. In order to validate the framework, the experimentation was conducted using two datasets from Twitter (one related to COVID-19 and the other related to the United States elections) in a distributed environment (Big Data). In both cases, the system considerably reduces the number of irrelevant users, just considering users with higher expertise.

There are also other approaches to address credibility-related problems. A framework for managing extracted knowledge from big social media data for decision-making is proposed in [24]. The framework allows extracting relevant knowledge from social media by comparing different social media knowledge. In [25], the authors present an ontology-based approach for identifying the credibility domain in

Social Big Data. They make use of ontologies in order to catch domain knowledge and enrich the semantics of the texts at both: entity and domain levels. Zhang et al. [26] develop a Scalable and Robust Truth Discovery (SRTD) scheme to tackle the misinformation spread and data sparsity challenges in social media in a distributed environment. The approach quantifies both the trustworthiness of sources and the credibility of claims.

As stated before, this research is an extended version of the work presented in [5]. Unlike previous approaches, our proposal provides a Big Data-based flexible system capable of generalizing and adapting to the user's needs. It also allows filtering content based on the user's experience. In this sense, the proposed system addresses two Big Data problems: the validity and volume of data.

Fig. 1 Reliable users for medical topics

3 Credibility Filtering Framework

Our credibility system is based on the fundamental premise that professionals often utilize their Twitter biographies to provide information about their occupations. We leverage this valuable resource to identify individuals who are particularly suitable as credible sources for specific use cases. To achieve this, we employ word embedding techniques to automatically assess the relevance and expertise of professionals based on the content of their biographies.

It is crucial to recognize that credibility depends on the alignment between an individual's professional background and the topic under consideration. For instance, an economist may have higher credibility when discussing stock market issues than a doctor. By leveraging word embedding techniques, our system can capture and analyze the semantic context within biographies, thereby enabling an automatic and objective assessment of credibility.

In this section, we comprehensively explain how we utilize word embedding models to process and evaluate the textual information contained within Twitter biographies. By outlining the technical aspects and methods employed, we aim to shed light on the robustness and reliability of our credibility assessment approach.

3.1 Twitter Biographies

Nowadays, social media has an irreplaceable role in our lives. It's a fact that millions of professionals use social media apps daily, like Twitter, to disseminate their work or to be informed. In [27], Oksa et al. analyzed social media usage from a work perspective, concluding that the usage of social networks for a professional is not only a fact but also brings improvements to the daily lives of workers in terms of networking, social support. According to [28], Twitter is one of the most widespread platforms to discuss recent advances or research in the field of medicine, but this success is also associated with more presence of misinformation or irrelevant content. Suppose we are browsing Twitter and we see some information that interests us. Probably the first thing we will do is to check that the account is reliable. For this, we will look at other data in the biographies since these are normally used to inform about jobs, likes or professions. So, in this paper, we propose an automatic system that exploits the potential of user biographies on Twitter and automatically discards accounts that do not fit with our desired analysis.

For example, suppose that we are interested in obtaining topics or clusters according to realistic medical information about COVID-19 from Twitter. For that, we need to obtain Tweets from people with some level of expertise in medicine. If we crawl Twitter data in order to apply a data mining algorithm, we will filter according to the hashtag #COVID19. Many users have tweeted about that topic using that hashtag, but only a few of them are interesting because they have expertise in the topic. In Fig. 1, we can see real accounts that have Tweeted about COVID that could be interesting for our topic. In contrast, in Fig. 2, we have people who tweeted about COVID in a trivial form. Our algorithm will select and discard those in a massive and automatic way. These figures are examples of some accounts that our algorithm considers relevant and others that our algorithm discards.

3.2 Expertise Filter

In this section, we go into detail about the expertise filter algorithm. In Fig. 3, a complete graph of the data flow through our algorithm can be seen. The algorithm takes, as input, the directory where we find the CSV files with the Tweets, a list of searches related to the topic under study and the

Fig. 2 Unreliable users for medical topics

language we are interested in. The dataset has been divided into small parts in order to apply an adaptation of the divide-and-conquer methodology.

The first step of the algorithm is a pre-processing module. In this stage, the cleaning of every Tweet is carried out. For this, the algorithm eliminates URLs, hashtags, mentions, reserved words from Twitter (RT, FAV...), emojis, smileys, numbers, additional spaces, and punctuation marks. Following this, all the textual terms are turned into lowercase letters. After this, the dataset language is detected, and according to that language, we obtain the related stopwords. After that, the system eliminates these stopwords and all those Tweets using a non-recognized language or another language different from the one desired by the user. Finally, any empty Tweet (composed of eliminated items in previous stages of pre-processing) is removed, and Tweet texts are tokenized.

Then, the core of the algorithm starts to operate. The functionality of the algorithm has been introduced using a use case. Let's imagine a case related to COVID-19, where experts or people related to science and medicine are wanted. The process starts with a list of words related to medicine introduced by the user; for example, *medical, doctor*. The algorithm will start to fit a word embedding model on the biographies of a part of a data partition (one of the input CSV files), and in the first iteration, it will obtain the most similar words to *medical* and *doctor* within the corpus itself according to a threshold. In the first version of our algorithm [5], the system obtained the five most similar words in each iteration, but we realized that this could lead to an exponential increase in the number of words obtained, which would also be of lower quality at each iteration. Also, the distance between the input and the selected words increases in each iteration, so the words become worse. To mitigate that problem, a similarity threshold has been introduced in the algorithm, so the user can set a value for the similarity between words and iterate over all the batches. If any word has a similarity higher than the threshold, then it is selected. It is worth emphasizing that we have employed word embeddings to capture the contextual information and to compute the similarity between words. When referring to similarity, we use the cosine similarity measure between the mean of the projection weight vectors for the words. We conducted a series of carefully designed experiments to determine the appropriate similarity threshold for word selection. We found that a threshold of 0.6 gave the most satisfactory results in terms of obtaining closely related words relevant to the given topic. By setting the threshold at 0.6, we struck a balance between retrieving words that have semantic associations with the topic and minimizing the inclusion of unrelated or less relevant terms. We tested various thresholds during our experimental phase and analyzed the outcomes. A threshold of 0.5 resulted in a higher number of unrelated words being included, which impacted the overall quality of the selected words. Conversely, a threshold of 0.7 or higher resulted in a significant reduction in the number of words, limiting the diversity and coverage of the selected terms. Therefore, we carefully selected the threshold of 0.6 to ensure an appropriate trade-off between precision and recall in word selection. It is crucial to bear in mind that a threshold value close to 1 implies a high degree of similarity, meaning that the selected word has an almost identical meaning or context to the topic. Conversely, a lower threshold implies a broader inclusion of words with varying degrees of similarity to the topic. By choosing the threshold of 0.6, we strike a balance between precision and inclusiveness, ensuring that the selected words are contextually relevant while avoiding an overly restrictive selection criterion.

Returning to our use case, in the first iteration, most similar words to *doctor* and *medical* were *itresearcher, medicine, researcher, physician, epidemic, pediatrician, epidemiologist, pediatrics, postdoctoral* and *toxicologist*. The algorithm will use these 12 words (the most similar to medical and the most 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, the retrieved words will be used to search for the most similar ones exceeding the threshold. This is one of the most flexible components of the system. Based on a very small set of one or two words, the algorithm automatically expands the query and flexibly locates words that could be very useful to filter the users according to their expertise in the related topic under study.

With the retrieved words, the algorithm selects the users who contain some of the words in their biographies, following the premise that biographies usually contain words

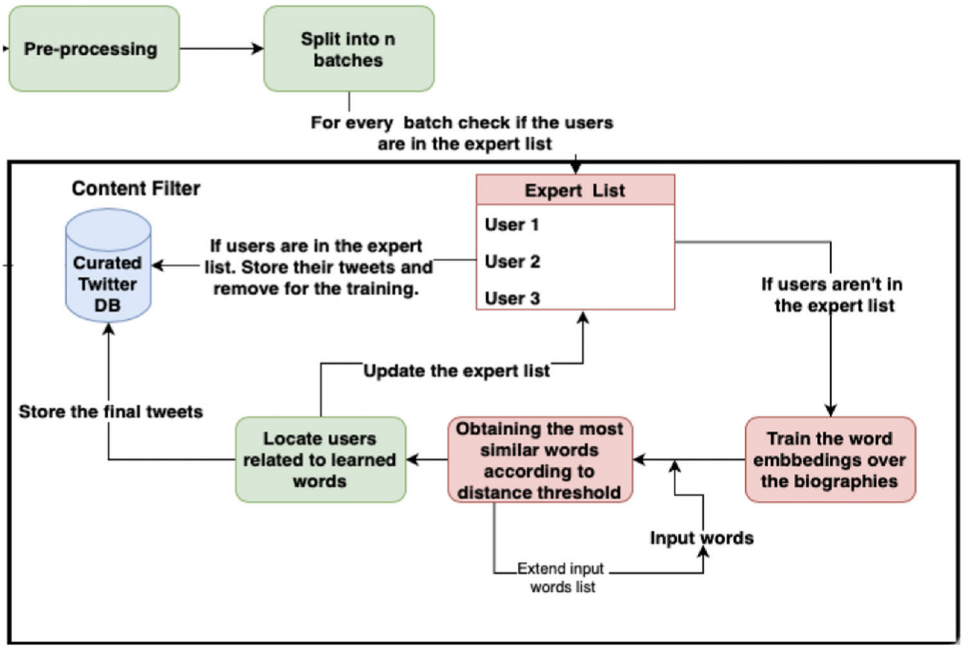

Fig. 3 Algorithm data flow

related to professions. The selected users are people with some level of expertise in the topic, so we called them experts. We store the experts in the expert list. 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 stages of the data mining pipeline. The complete pseudocode of the algorithm is in Algorithm 1.

4 Big Data Architecture

With the increasing size of the data generated and stored, traditional Data Mining and data pre-processing techniques face a great challenge in efficiently processing large data sets. For this reason, the use of distributed computing has been used as a solution even before the Big Data phenomenon. This type of massive data processing does not only require adapting existing algorithms but also proposing new ones to handle Big Data problems.

MapReduce (MR) is one of the first distributed computing paradigms that allowed the generation and processing of Big Data datasets in an automatic and distributed way. MR has become the benchmark in distributed computing paradigms because of its simplicity and fault tolerance. By implementing two functions, Map and $Reduce$, users can process large amounts of data without worrying about technical issues such as data partitioning, fault recovery, or job communication. The algorithm we propose has been designed to enhance and use the full capacity of any data cluster using Spark. For this purpose, it has been implemented using several processes in which we use MapReduce.

To design the distributed algorithm for the credibility filter, some primitive Spark functions would be necessary. It is explained here:

- Map: Applies a transformation function to each element of RDD and returns a transformed RDD.
- $FlatMap$: Similar to Map, but each input item can be mapped to 0 or more output items.
- $Reduce$: Aggregates the elements of the dataset using an aggregation function.

Additionally, the algorithm uses broadcast variables to enable access to global variables in every node of the cluster, i.e., broadcast variables are available in every partition performed by the Map functions. The shared variables are stored in a hash table format, allowing direct and fast access to the query and insertion of the expert.

The algorithm has two main steps (see Algorithm 1). The first one consists of loading the dataset and filtering the experts from each dataset using the FlatMap function. The second step consists of aggregating all these data using a reduce function and grouping them by each found expert.

In the first stage of the algorithm, it needs the users and tweets with the information processed by the *FindExpert()* function. For this purpose, a *FlatMap* function is used (see line 5 of the Algorithm 1 and the first column of Fig. 4).

The *FindExpert* function is described in Algorithm 2, which filters the *expert list* and returns the important informa-

Algorithm 1 Main Spark procedure for expertise filter algorithm

1: **Input:** *Data:* RDD transactions: $\{t_1, \ldots, t_n\}$
2: **Input:** *similarity_threshold:* Similarity between words
3: **Output:** $Global\_expert\_set$: Expert discover in each $FlatMap$
4: Distributive computing in q chunks of transactions: $\{S_1, \ldots, S_q\}$
5: $\{< Expert\_1 > \ldots < Expert\_m >\} \leftarrow S_i$.**FlatMap**(FindExpert())
6: $FinalDataframe \leftarrow$ **ReduceByKey** $(getInfo())$
7: End distributive computation
8: **return** $FinalDataframe$

Algorithm 2 FindExpert

1: **Input:** *datainput:* Tweet data and topics
2: **Input:** *similarity_threshold:* $minSim \in (0, 1]$
3: **Output:** *KeyValuePair:* Topic and Expert
4: **while** $ExpertList \neq null$ **do**
5: $Expert \leftarrow ExpertList.pop()$
6: **if** $ExpertID \in Global\_expert\_set$ **then**
7: # For each expert, we add all their tweets to the final data frame
8: $Expert_{final} \leftarrow < Expert_{id}, tweets >$
9: $output.append(Expert_{final})$
10: **else**
11: # Finding new experts by processing the rest of the content
12: $tokenized\_tweet = data['clean\_bios']$
13: $model = train(tokenized\_tweet)$
14: # Adding the most similar words in the model
15: $most\_similar = SimilarFinalWords()$
16: $expert\_words.extend(most\_similar)$
17: # Finding users having any of the words in their biographies
18: $output.append(< Expert_{id}, [expert\_words, clean\_bios, tweets] >)$
19: **end if**
20: **end while**
21: **return** $Output$

Algorithm 3 SimilarFinalWords

1: **Input:** *expert_words:* Words from each expert
2: **Input:** *similarity_threshold:* $minSim \in (0, 1]$
3: **Output:** $KeyValuePair$: Topic and Expert
4: **while** $wordinexpert\_words$ **do**
5: **if** $wordinmodel$ **then**
6: $final\_words.append(word)$
7: **end if**
8: **end while**
9: # Data frame creation with words most similar to each expert word according to the similarity threshold
10: **while** $wordinfinal\_words$ **do**
11: **if** $word.similarity > minSim$ **then**
12: $most\_similar.append(find\_most\_similar(model, word))$
13: **end if**
14: **end while**
15: **return** $most\_similar$

tion for each of them. The function starts by training a word embedding model on the biographies of the data partition found in the first iteration. To do this, the five most similar words to those that are passed as parameters (e.g., democratic, republican) are found. Algorithm 2 is responsible for aggregating similar words between the experts described in Algorithm 3. The algorithm will use these 12 words (five most similar to democratic, five most similar to republican, in addition to democratic and republican) to find users whose biographies contain any of these terms and start creating the list of experts that will be stored in the distributed variable $Global\_expert\_set$ (line 3 of Algorithm 1). In the next iteration, these 12 words will be used to search for their five similar ones, and so on.

The FindExpert function will return a list of peers containing the expert and its expert words and some other information. To aggregate and obtain all the information generated in a distributed manner, a *Reduce* function (see line 6 of Algorithm 1 and Fig. 4) is used. As a result, a set of words associated with each expert is obtained.

To understand the computational complexity of our proposal, we can see how the pseudocode performs a search for tweets. For each of these searches, we search the experts, so the complexity of the algorithm is $M * N$ where M is the number of tweets and N is the number of experts found. So in the sequential case, the complexity would be $O(M * N)$ and $O(\frac{M*N}{K})$, where K is the number of data partitions executed concurrently.

5 Experiments

In this section, the experimentation has been carried out to analyze the system efficiency on three different use cases (sport, health and politics) in detail. The interpretation of the results will be discussed in Sect. 6.

5.1 Dataset

The system has been performed on three different real datasets without any processing of the social network Twitter according to the three mentioned use cases. To accomplish this, we utilized the Twitter streaming API without keywords, applying language filtering to English content and restricting the timeframe to the period between 5th to 10th August 2022. Taking into account the restrictions of the API in terms of time windows and that the data collection script was used intermittently during these days to avoid biases of topics that monopolized the network on a single day, the total number of tweets obtained is 5,000,000 in 20 different splits of 250,000 tweets.

The content of the dataset is not filtered in any way, i.e., it contains any tweet that was tweeted in the time windows in which the data were collected. Therefore, it can be concluded that it is a noisy and very low-quality dataset. When considering datasets within the big data paradigm, it is important to note that the majority of the articles reviewed in Sect. 2.2 consist of a relatively small number of total tweets compared to our study. However, there is one notable exception: the

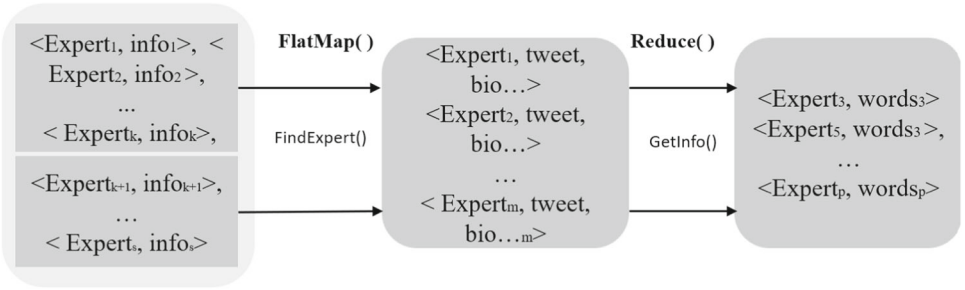

Fig. 4 Workflow of Big Data process in Spark

CredSat system [15]. This system employs a similar volume of tweets for analysis, namely 5,220,478 tweets. This makes it one of the few credibility and expertise systems, along with ours, to use a significant number of tweets in its dataset. The inclusion of a large number of tweets in our research is essential. It allows us to comprehensively analyze and understand credibility and expertise, as a larger dataset provides a broader representation of user behavior and content. This allows us to make more robust assessments and derive meaningful insights from the data, rather than other papers that are framed within big data and use a few examples. Finally, the last reason for framing this paper within the big data paradigm is that it is fully operational. As we can see in Fig. 5, the time to process 5 million tweets is unique, so in the Big Data paradigm with Spark, our filter is a fully operational and functional solution.

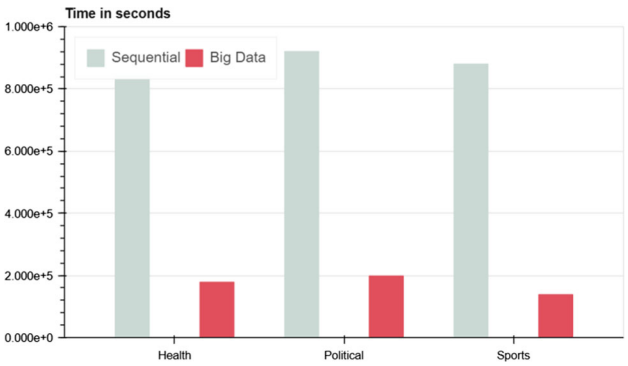

Fig. 5 Execution time comparison for 5 M tweets and each use case

Table 2 Machine specifications

Component	Features
CPU	4 x 3 GHz Intel Xeon E5 with 8 cores
RAM	240 GB 3200 MHz LPDDR4X
Hard Disk	SATA SSD de 6 TB

5.2 Results

With the objective of validating the results and the improvement in terms of efficiency of the proposed big data system, in this section, comparative experimentation has been carried out between the sequential execution of the algorithm and the execution according to the big data paradigm proposed in Sect. 4. The technical specifications of the computer on which our experiments were run can be found in Table 2.

The first version of the paper [5] has concluded that the best configuration in terms of a ratio between expert located and time was Fast-Text [29] + Skip Gram. So, for the final version of our algorithm, this configuration has been chosen. Regarding the embedding parameters, it has been run with a window of 5 words. Words with frequencies lower than two have been ignored, and also having hierarchical softmax. The results using the mentioned seen configurations, for a mean score of 5 different executions for each use case, can be found in Table 3.

This table shows the experimental results on the entire dataset. As can be seen, the improvement in time of the sequential version versus the distributed version is remarkable. It is discussed in more detail by analyzing the performance and efficiency. As for the localized experts, it can be seen how the proposed system can extract a subset by performing a 90% reduction of the dataset.

Figure 6 shows the memory usage of each algorithm for every use case. It can be observed that the distributed algorithm does not outperform the sequential case in all cases. This is due to the treatment of the data as they are replicated in different machines, which duplicates information.

With the purpose of measuring the efficiency of our proposal and comparing it to the existent approaches, it has been analyzed the *speed up* and the *efficiency* [30, 31] according to the number of cores. For that, we have computed the well-known measure of speed up defined as [32]:

$$S_n = T_1/T_n, \tag{1}$$

where T_1 is the time of the sequential algorithm, and T_n is the execution time of the distributed algorithm using several cores. The efficiency [30–32] is defined in a similar way as:

$$E_n = S_n/n = T_1/(n \cdot T_n). \tag{2}$$

Table 3 Comparison of results between sequential and big data executions

Use case	Time	Experts located	Final dataset size	% Reduction
Health	842,350	72,290	97,342	98.05
Political	920,235	98,228	110,293	97.7
Sports	879,817	123,342	185,323	96.29
BigData-Health	178,320	76,170	89,746	98.20
BigData-Political	198,320	90,674	137,827	97.24
BigData-Sports	138,320	110,493	167,425	96.65

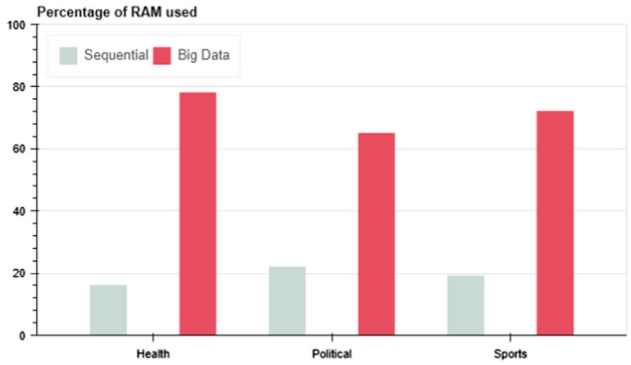

Fig. 6 Percentage of memory used for the entire dataset

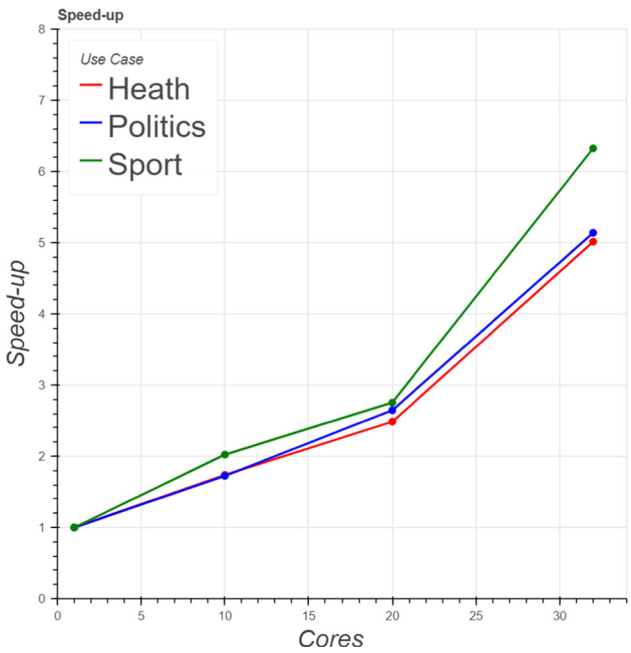

Fig. 7 Speed-up of different use cases for 2 million of records

Figures 7 and 8 show the results obtained for the database with 2 million tweets. In them, it can be seen that the efficiency and speed improve as the number of cores increases, although they are not optimal. This is due to the workloads of the cores and the congestion of the network used for network communication between the cores. In addition, Fig. 7 shows that the speed-up increases over the number of processors

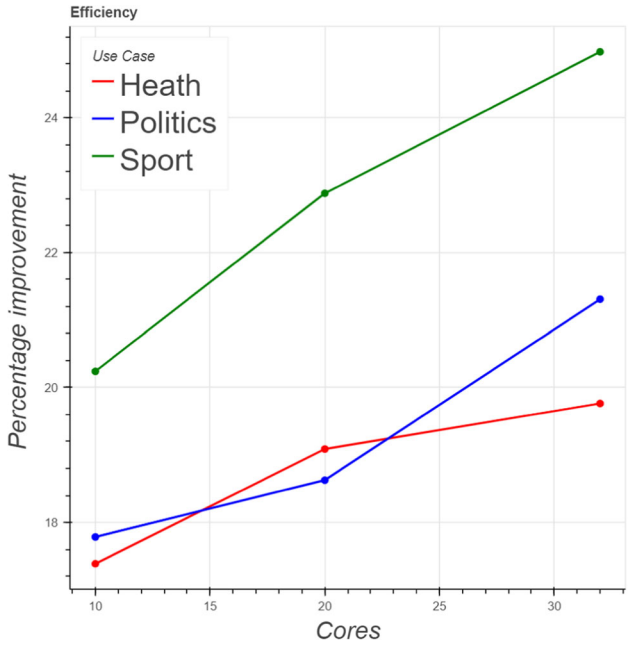

Fig. 8 Efficiency of different use cases measured by the percentage of improvement for 2 million of records

used, although it is not proportional as the resources increase (see also how the efficiency does not increase in Fig. 8). This same behavior can be observed in other studies of speed up and efficiency in distributed algorithms, where the efficiency is not improved in a proportional way, as desired, with more processing cores [32].

Another advantage of the proposed algorithm is that it is based on a technology that allows the use of large data clusters in a simple way. Thanks to this, Larger clusters or cloud computing systems such as AWS or Google Cloud can be used if we need more capacity to process larger data sets.

In the results obtained, high-quality expert users on a specific topic have been identified using Big Data. This process is highly relevant in this context because one of the characteristics of Big Data is the volume of data. Using this filter, higher-quality data can be validated within massive datasets. In the following section, we will analyze and validate some of the results obtained.

Table 4 Learned words for each use case

Use case	Input words	Top 20 earned words
Health	Doctor, researcher	Research, doctor, surgeon, researcher, medicine, medical, lecturer, clinical, physician, epidemic, paediatrics, epidemiologist, exclinical, postdoctoral, toxicologist, biologist, physiologist, cardiology, aesthetician, virologist
Sports	Football, baseball	Football, baseball, softball, redsox, hockey, basketball, football, sportsnet, volleyball, rugby, jockey, celtic, coach, intense, whitesox, paintball, handbasket, cricket, lebron, sport
Political	Democratic, republican	Democrat, socialist, trump, democratic, secular, republican, protrump, repubs, liberal, conservative, humanist, socialism, joebiden, exlabour, ecosocialist, freethinker, autocratic, exrepublican, demsocialist, conservatism

6 System Validation

In order to validate the proper functionality of the system, we have applied the algorithm to our dataset with three different sets of input words. Each set of input words corresponds with a different use case, one related to health, another to sports and another to politics. These different use cases demonstrate that the system is flexible and valid in multiple domains. Also, we demonstrate that our algorithm is suitable, in a very flexible way, to transform a non-quality and noisy dataset into a high-quality dataset. In Table 4, the input words and the top 20 learned words for each use case can be seen. At this point, we can derive one of the major contributions of our paper on the topic of flexible systems. If we pay attention to the learned words, it is easy to see how the algorithm retrieves closely related words to each domain in a semi-supervised way, only using a pair of input words. The system adjusts perfectly to each domain in a simple and effective way for the user who sees how his query is flexibly expanded according to the domain and with little effort on the user's part since only one or two words are needed as input.

Also, the proposed algorithm is able to mitigate one of the traditional problems in user-generated content, misspelling. Twitter biographies are user-generated text and usually contain typos or domain-related user-generated words, i.e., hashtags. Our algorithm can link those words with the topic. For example, in the case of sports, some interesting words that the algorithm used to filter the content are *fuball* or *fotball*, two misspelling words of *football*. Also, in the case of politics, we find *conserv* or *repub*, misspelling words or *resistbiden* and *jailtrump*, two user-generated words for political discussions in Twitter. That result led us to argue that the system is flexible and robust.

Using the set of words learned for each use case, the algorithm can filter the Twitter accounts and select only those users that fit with the topic. With that, from a non-quality and hyper-generic dataset, we can obtain valuable information. In order to prove that, we obtain word clouds over the filtered datasets. We try to prove the premise that if the system performs properly, visualization techniques over the filtered dataset must obtain closely related topic words. We must bear in mind that from this moment on, we always refer to the textual content of tweets. The algorithm has used the biographies to obtain the most reliable accounts according to their expertise on the topic. Now, we only focus on the content of the Tweet, trying to demonstrate that the filtering process proposed in this paper improves the quality of the dataset.

It is important to highlight that a named entity recognition process was applied to the filtered datasets for each specific use case. Due to the specificity of the medical domain for the health use case, we have used the *ner_bionlp13cg_md* model included in SciSpacy [33]. This model was trained on the BIONLP13CG corpus. Regarding the political and sports, which are use cases with more general domains, we have used the *en_core_web_sm* included in SpaCy [34]. This was trained using web blogs, news and comments. Word clouds were then created using two different text representations over the named entities, one based on TF-IDF [35] and the other one based on traditional term frequency. Word clouds were used to visually represent and highlight the most significant words for each topic and use case, particularly in the areas of sport, politics, and health. Word clouds have been widely adopted as a popular technique for topic representation, as they provide a simple and intuitive way to gain insight into the most prominent words within a given corpus. It is important to note that other visualization techniques and methods of topic explainability were considered but ultimately discarded for this particular analysis. This decision was made because the primary objective of the study was to demonstrate how the application of the credibility filter significantly enhances the quality of the dataset by excluding irrelevant information that is not related to the specific use case or topic under analysis. By focusing on word clouds, the analysis can effectively demonstrate the impact of the credibility filter in improving data quality and relevance.

If we pay attention to the word clouds (Figs. 9, 10, 11, 12, 13, 14), we can see how we can easily locate topic-related words in each example. There are a few examples that appear in all the figures. If we take into account that all the results

Fig. 9 Word cloud for the health use-case using the frequency of words

Fig. 10 Word cloud for the health use-case using the TF-IDF value

Fig. 11 Word cloud for the political use-case using the frequency of words

Fig. 12 Word cloud for the political use-case using the TF-IDF value

come from the same dataset, the low value of common words and the topic-related words present in each use case led us to argue that our filter works in a very proper way and it can discard noise and irrelevant information from a big data set. Also, some words can be present in all the domains, but this is due to the bias of the data acquisition stage, and because these words like, for instance, *people*, or *american* or some verbs are very common in Twitter's conversations.

In the case of health, Figs. 9 and 12 are very similar due to the specification of the NER process in this topic. In the other use cases, the frequency and TF-IDF figures are too different, and maybe in the first look at Fig. 11 and Fig. 13, we can think that the algorithm does not work properly. In a deeper analysis we browsed the social network Twitter trying to find the meaning of that words, and we realized that they were usernames. The vast majority of these user names come from users related to sports brands, basket players and people related to sports in the case of sports. And users who are politicians, companies or journalists in the case of politics. In Fig. 15 and, we can see the most representative users located in the conversations of Twitter for each use case.

Considering that the dataset was obtained without any filter having as results data sets in which we can easily obtain closely related topic information and find people that usually participate in the conversations about each topic is a very

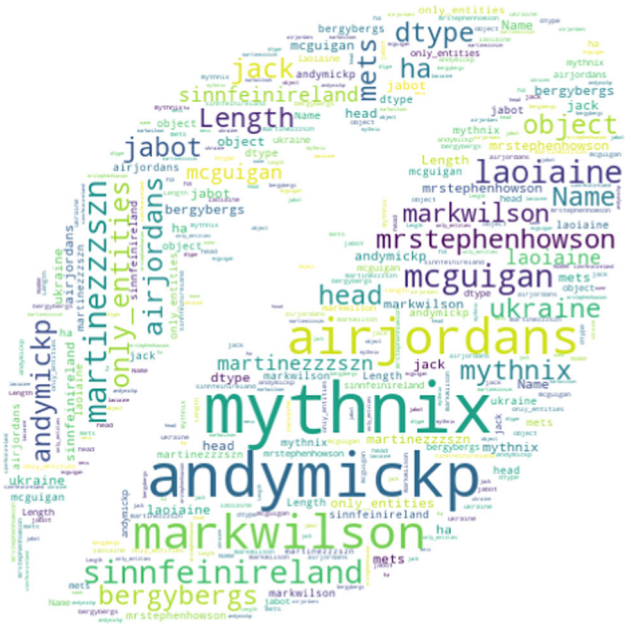

Fig. 13 Word cloud for the sports use-case using the frequency of words

Fig. 14 Word cloud for the sports use-case using the TF-IDF value

important result. It is interesting to note how some accounts located in the Twitter conversations by the entire pipeline do not have any information in the user names or biographies about the topic, but yes in the images or main user image. This again led us to argue the potential of the framework to locate interesting information for hundreds of experts in a flexible and non-supervised way. Based on these findings, we can conclude that our filter works properly and can get value and information regarding interesting topic-related conversations and users from a non-quality and noisy dataset.

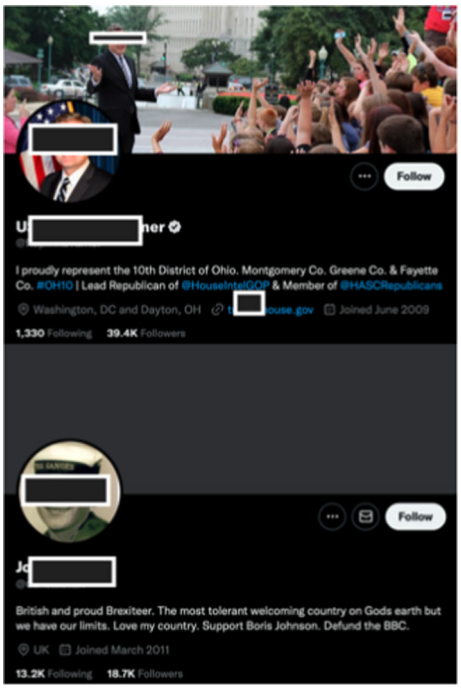

Fig. 15 Some users mined by the filter and the pipeline for the political use-case

7 Conclusions and Future Work

In this paper, we have proposed a flexible, easy-to-use, easy-to-understand and easy-to-replicate big data system for filtering the content coming from Twitter based on the credibility of the user account based on his or her expertise in a certain topic. We have undertaken comprehensive experimentation to validate the efficacy of the big data architecture proposed in this paper, demonstrating its capability to significantly enhance processing time compared to traditional methods. We also have conducted three different use cases, in which we have demonstrated the following:

- The proposed system can help to acquire more valuable datasets in a flexible and unsupervised way.
- The proposed system can help to reduce the dimensionality of a big data set, maintaining only the interesting examples for the desired use case.

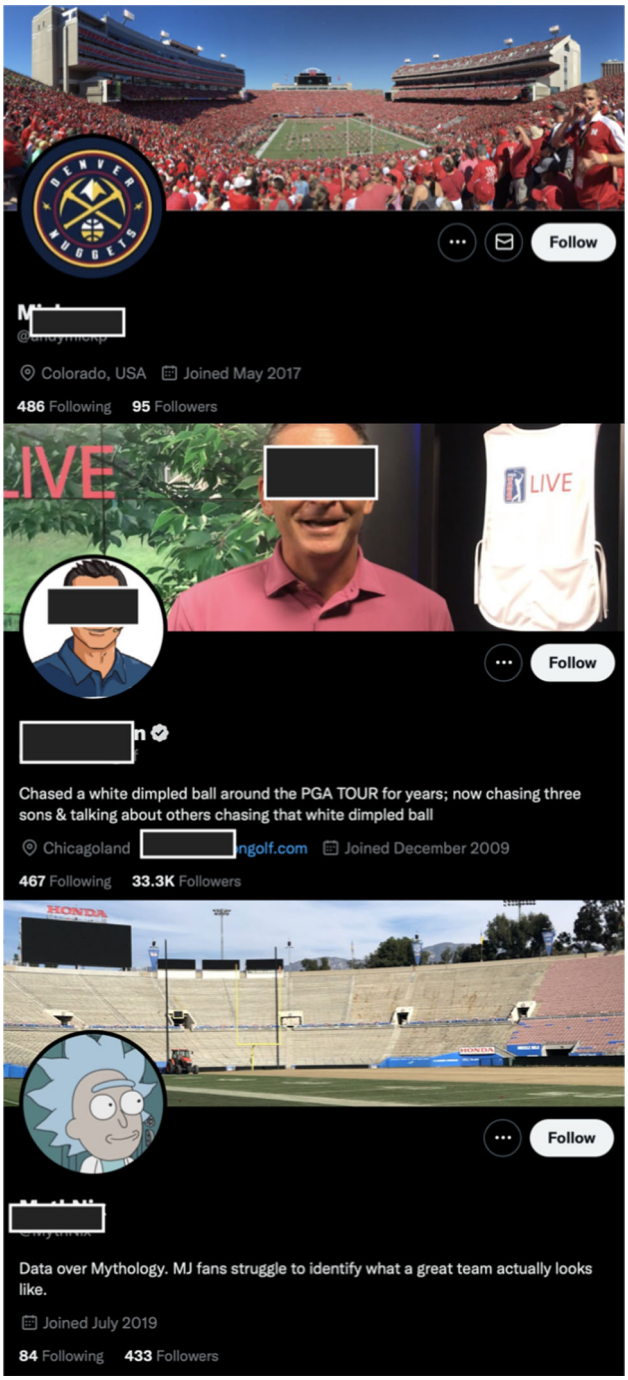

Fig. 16 Some users mined by the filter and the pipeline for the sports use-case

- The proposed system can help to improve the validity of the data in terms of expertise.

The system is sensitive to bias in data acquisition. Sometimes time windows can be more interesting than other windows, depending on which moment we collect the data to obtain better o worse results. In this paper, we wanted to demonstrate how our algorithm can obtain a curated dataset over a noisy dataset, so we created a real-word dataset using time windows without content filters. But, real word applications, usually the data acquisition, are made with filters, for example, using hashtags, so these applications are independent of the time windows. In these problems, our system can also improve the subsequent data mining process [16]. In these real-world applications with filters in data capture, the systems obtain even better results, discarding people without expertise in the topic and maintaining very useful accounts in which we can perform more accurate analysis. Regarding the limitations of the system, since the core of the system (Algorithm 1) is based on Twitter biographies, it is very sensitive to lies in biographies.

In terms of future work, there is potential to adapt the system into a streaming application that can be incorporated into an incremental learning pipeline. This would enable real-time filtering of irrelevant content for the specific topic under study, resulting in more accurate and credible results. Such an application would be particularly valuable in addressing the challenge of misinformation spread on social networks, such as Twitter or Reddit. Furthermore, it is crucial to consider the emerging trend of discarding misinformation as part of the information retrieval process. This domain, where our research and other pioneering papers [20, 22] are situated, holds promise for addressing common challenges associated with misinformation, including enhancing the generalization capabilities of systems. Consequently, continual improvement and development of systems in this field are imperative.

Acknowledgements The research reported in this paper was supported by the FederaMed project: Grant PID2021-123960OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU and DesinfoScan project: Grant TED2021-129402B-C21 funded by MCIN/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309). 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).

Author Contributions Jose A Diaz-Garcia: supervision, investigation, project administration, writing—original draft and editing. Carlos Fernandez-Basso: investigation, software, writing—original draft. Karel Guitierrez-Batista: investigation, software, writing—original draft. M. Dolores Ruiz: funding acquisition, project administration, writing—review and editing. Maria J. Martin-Bautista: funding acquisition, project administration, writing—review and editing.

Funding The research presented in this paper has received funding from: FederaMed project: Grant PID2021-123960OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU and the BIGDATAMED projects with references B-TIC-145-UGR18 and P18-RT-2947. The European Union NextGenerationEU /PRTR, grant PLEC2021-007681 funded by MCIN/AEI/10.13039/501100011033. The DESINFOSCAN project. Ministerio de Ciencia e Innovacion and by the European Union NextGenerationEU (Grant TED2021-1289402B-C21). The NOFACEPS project (PPJIB2021-04) of the

University of Granada's internal plan. Carlos Fernandez-Basso was supported by the Ministry of Universities through the EU-funded Margarita Salas Programme. Jose A. Diaz-Garcia was supported by the Spanish Ministry of Education, Culture and Sport (FPU18/00150). Karel Gutiérrez-Batista was supported by the Administration of the Junta de Andalucía.

Availability of data and material Data will be available on request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval No ethical approval is required for this study.

Consent to participate Not applicable.

References

1. Perrin, A.: Social media usage. Pew Res. Center **125**, 52–68 (2015)
2. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. Ai & Society **30**(1), 89–116 (2015)
3. Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., Yu, Z.: Text mining of user-generated content (ugc) for business applications in e-commerce: a systematic review. Mathematics **10**(19) (2022). https://doi.org/10.3390/math10193554
4. Assefi, M., Behravesh, E., Liu, G., Tafti, A.P.: Big data machine learning using apache spark mllib. IEEE Int. Conf. Big Data (Big Data) **2017**, 3492–3498 (2017). https://doi.org/10.1109/BigData.2017.8258338
5. Diaz-Garcia, J. A., Ruiz M. D., Martin-Bautista, M. J.: A comparative study of word embeddings for the construction of a social media expert filter. In: International Conference on Flexible Query Answering Systems. Springer, 196–208 (2021)
6. Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I.: Apache spark: a unified engine for big data processing. Commun. ACM **59**(11), 56–65 (2016). https://doi.org/10.1145/2934664
7. Fernandez-Basso, C., Ruiz, M. D., Martin-Bautista, M. J.: New spark solutions for distributed frequent itemset and association rule mining algorithms. Cluster Comput., 1–18 (2023)
8. Honnibal, M., Montani, I.: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, Unpublished software application (2017). https://spacy.io
9. Sharnagat, R.: Named entity recognition: a literature survey. Center For Indian Language Technology, 1–27 (2014)
10. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 675–684 (2011)
11. Canini, K. R., Suh, B., Pirolli, P. L.: 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, IEEE, pp. 1–8 (2011)
12. Hassan, D.: A text mining approach for evaluating event credibility on twitter. In: 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), IEEE, 171–174 (2018)
13. Alrubaian, M., Al-Qurishi, M., Hassan, M.M., Alamri, A.: A credibility analysis system for assessing information on twitter. IEEE Trans. Depend. Secure Comput. **15**(4), 661–674 (2016)
14. Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M.M., Alamri, A.: Reputation-based credibility analysis of twitter social network users. Concurr. Comput. Pract. Exp. **29**(7), e3873 (2017)
15. Abu-Salih, B., Wongthongtham, P., Chan, K.Y., Zhu, D.: Credsat: credibility ranking of users in big social data incorporating semantic analysis and temporal factor. J. Inform. Sci. **45**(2), 259–280 (2019)
16. Diaz-Garcia, J. A., Ruiz, M. D., Martin-Bautista, M. J.: Noface: A new framework for irrelevant content filtering in social media according to credibility and expertise. Expert Syst. Appl. 118063 (2022)
17. Cordeiro, P. R. D., Pinheiro, V., Moreira, R., Carvalho, C., Freire, L.: What is real or fake?-Machine learning approaches for rumor verification using stance classification. In: IEEE/WIC/ACM International Conference on Web Intelligence, 429–432 (2019)
18. Ozbay, F.A., Alatas, B.: Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Stat. Mech. Appl. **540**, 123174 (2020)
19. Batra, J., Jain, R., Tikkiwal, V.A., Chakraborty, A.: A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. Int. J. Inform. Manag. Data Insights **1**(1), 100006 (2021)
20. Pradeep, R., Ma, X., Nogueira, R., Lin, J.: Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2066–2070 (2021)
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
22. Zhang, D., Vakili Tahami, A., Abualsaud, M., Smucker, M. D.: Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2099–2104 (2022)
23. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information-a survey. Wiley Interdiscip. Rev. Data Mining Knowl. Discov. **7**(5), e1209 (2017)
24. He, W., Wang, F.-K., Akula, V.: Managing extracted knowledge from big social media data for business decision making. J. Knowl. Manag. (2017)
25. Wongthongtham, P., Salih, B.A.: Ontology-based approach for identifying the credibility domain in social big data. J. Organ. Comput. Electron. Comm. **28**(4), 354–377 (2018)
26. Zhang, D., Wang, D., Vance, N., Zhang, Y., Mike, S.: On scalable and robust truth discovery in big data social media sensing applications. IEEE Trans. Big Data **5**(2), 195–208 (2018)
27. Oksa, R., Kaakinen, M., Savela, N., Ellonen, N., Oksanen, A.: Professional social media usage: work engagement perspective. New Media Soc. **23**(8), 2303–2326 (2021)

28. Pershad, Y., Hangge, P.T., Albadawi, H., Oklu, R.: Social medicine: Twitter in healthcare. J. Clin. Med. **7**(6), 121 (2018)
29. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
30. Kumar, V.P., Gupta, A.: Analyzing scalability of parallel algorithms and architectures. J. parall. Distrib. Comput. **22**(3), 379–391 (1994)
31. Grama, A.Y., Gupta, A., Kumar, V.: Isoefficiency: measuring the scalability of parallel algorithms and architectures. IEEE Parall. Distrib. Technol. Syst. Appl. **1**(3), 12–21 (1993)
32. Barba-González, C., García-Nieto, J., Benítez-Hidalgo, A., Nebro, A.J., Aldana-Montes, J.F.: Scalable inference of gene regulatory networks with the spark distributed computing platform. In: Del Ser, J., Osaba, E., Bilbao, M.N., Sanchez-Medina, J.J., Vecchio, M., Yang, X.-S. (eds.) Intelligent Distributed Computing XII, pp. 61–70. Springer International Publishing, Cham (2018)
33. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, (2019), 319–327. https://doi.org/10.18653/v1/W19-5034
34. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spacy: Industrial-strength natural language processing in python (2020). https://doi.org/10.5281/zenodo.1212303
35. Qaiser, S., Ali, R.: Text mining: use of tf-idf to examine the relevance of words to documents. Int. J. Comput. Appl. **181**(1), 25–29 (2018)