applied sciences

MDPI

Article

Credibility Analysis on Twitter Considering Topic Detection

Maria Hernandez-Mendoza [1], Ana Aguilera [2,*], Irvin Dongo [3,4], Jose Cornejo-Lupa [5] and Yudith Cardinale [1,3,6,*]

[1] Departamento de Computación y Tecnología de la Información, Universidad Simón Bolívar, Caracas 1080, Venezuela
[2] Escuela de Ingeniería Informática, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso 2340000, Chile
[3] Electrical and Electronics Engineering Department, Universidad Católica San Pablo, Arequipa 04001, Peru
[4] ESTIA Institute of Technology, University of Bordeaux, 64210 Bidart, France
[5] Computer Science Department, Universidad Católica San Pablo, Arequipa 04001, Peru
[6] Escuela Superior de Ingeniería, Ciencia y Tecnología, Universidad Internacional de Valencia, 46002 Valencia, Spain
* Correspondence: ana.aguilera@uv.cl (A.A.); yudith.cardinale@campusviu.es (Y.C.); Tel.: +56-322603735 (A.A.); +34-623423734 (Y.C.)

Abstract: Twitter is one of the most popular sources of information available on the internet. Thus, many studies have proposed tools and models to analyze the credibility of the information shared. The credibility analysis on Twitter is generally supported by measures that consider the text, the user, and the social impact of text and user. More recently, identifying the topic of tweets is becoming an interesting aspect for many applications that analyze Twitter as a source of information, for example, to detect trends, to filter or classify tweets, to identify fake news, or even to measure a tweet's credibility. In most of these cases, the hashtags represent important elements to consider to identify the topics. In a previous work, we presented a credibility model based on text, user, and social credibility measures, and a framework called T-CREo, implemented as an extension of Google Chrome. In this paper, we propose an extension of our previous credibility model by integrating the detection of the topic in the tweet and calculating the topic credibility measure by considering hashtags. To do so, we evaluate and compare different topic detection algorithms, to finally integrate in our framework T-CREo, the one with better results. To evaluate the performance improvement of our extended credibility model and show the impact of hashtags, we performed experiments in the context of fake news detection using the PHEME dataset. Results demonstrate an improvement in our extended credibility model with respect to the original one, with up to 3.04% F1 score when applying our approach to the whole PHEME dataset and up to 9.60% F1 score when only considering tweets that contain hashtags from PHEME dataset, demonstrating the impact of hashtags in the topic detection process.

Keywords: credibility model; topic detection; Twitter

check for updates

Citation: Hernandez-Mendoza, M.; Aguilera, A.; Dongo, I.; Cornejo-Lupa, J.; Cardinale, Y. Credibility Analysis on Twitter Considering Topic Detection. *Appl. Sci.* **2022**, *12*, 9081. https://doi.org/10.3390/app12189081

Academic Editor: Valentino Santucci

Received: 8 August 2022
Accepted: 6 September 2022
Published: 9 September 2022

1. Introduction

Social networks have become tools in people's daily life to share, for example, their opinions, feelings, and stories [1], as well as to support their professional life to communicate news, disasters, accidents, etc. [2]. Thus, social media contributes significantly to a variety of situations, such as awareness [3], disaster notifications [4], entertainment, communication, news and social interaction, information sharing, information seeking, self-documentation, and self-expression [5].

Among the current social media platforms, Twitter is one of the more widely used throughout the world [6], having 650 million registered users [7]. It is the largest social network used to write and read people's short text (called tweets) about anything in life, with a maximum of 280 characters, mixed with contextual clues, such as URLs, tags, usernames and informal misspelling, acronyms and abbreviations, and can also contain videos and photos [8]. By default, all Twitter accounts are public; thus, anyone can read the tweets published in any account. The connected network of users is built in terms of *followed* (i.e., accounts that a user chooses to follow) and *followers* (i.e., users that follow an account); a user's "timeline" includes chronological updates of tweets from the users they follow [7]. Each user has a social influence on his/her network and can make mentions of other tweets or replies of tweets (retweets).

The popularity of Twitter is evident [9], and its success is partly due to the facility of information access for the masses and cheap cost [10,11]. Twitter also uses a very effective and scalable infrastructure to implement a straightforward data delivery paradigm [12]. However, similar to other social networks, Twitter has been the victim of several hostile attempts. In general, online social networks have the potential to be misused to spread false information, engage in political censorship, sway public opinion, and manipulate users [13]. Huge quantities of fake news [14–16], rumors [17], hoaxes [18], and trending news [19] are disseminated daily, which have the potential for extremely negative impacts on individuals and society [20,21].

Unfortunately, as an information source, Twitter—as well as other social media platforms—has neither a technique nor ranking for veracity to allow inferring the credibility level of information. The verification of information on these social media platforms is absent [22]. In the literature, several works have proposed credibility models for Twitter analysis based on three measures: text credibility, user credibility, and text and user social credibility [22–26]. Some other works have also considered topic detection in the context of Twitter credibility [3,27–29]. However, most of the works use topic detection as a technique for discovering trends [30], discovering natural disasters as early as possible [31], helping political parties and companies to understand users opinions [32], improving content marketing by better understanding customer needs, and even more [33].

Thus, there is still a need for further research in this area, which is complicated by the characteristics of tweets: short texts, large volume of tweets, noisy and unstructured data [34]. As tweets are short texts, they usually include misspelled words, irrelevant characters, emojis, unconventional syntax, and hashtags, among other elements, which can negatively or positively affect the performance of topic detection algorithms [35,36].

Many works have studied the value and significance of hashtags and how to help users select hashtags more efficiently (hashtag recommendation)—that is, a hashtag related to the topic of the tweet [37,38]. Therefore, the use of hashtags can have a high impact on the topic used in the tweet, since they are a word or concatenation of short words that are usually used to define the topic that is being talked about, without the need to read the full tweet [39]. It is usually used so that the tweet has a greater reach if another person is looking for tweets that contain the same hashtag. There are also studies that discuss how topic detection can be improved if the hashtag is included as part of the model [35,40].

In this context, we pretend to combine the insights about topic detection and hashtag usage to ameliorate a Twitter credibility model. In previous works, we proposed a credibility model [22] and a framework called T-CREo [41] to evaluate Twitter credibility. Our previous model considers text, user, and social credibility measures to calculate the overall credibility level of tweets. This model has been integrated into the T-CREo framework implemented as an extension of Google Chrome allowing the real time analysis of tweets [41]. In this work, we propose an extension of our previous credibility model by incorporating the topic analysis measure based on the evaluation of hashtags. The topic that people talk about on a tweet is identified and used as an additional metric in the model for calculating the tweet's credibility. Thus, the proposed extended model considers four levels of credibility (i.e., text, user, social, and topic), each one contributing 25% (by default) of the tweet's entire credibility.

To obtain the topic credibility measure, we evaluate and compare several topic detection algorithms to integrate the best one into the T-CREo framework [41]. To evaluate the performance improvement of our extended credibility model and its implementation in

T-CREo, we performed experiments in the context of fake news detection using the PHEME dataset. Results show an improvement in the extended credibility model with respect to the original, i.e., up to 3.04% F1 score when applying our approach to the whole PHEME dataset and up to 9.60% F1 score when only considering tweets that contain hashtags.

In summary, the main contributions of this work are as follows: (i) The integration, into a previous credibility model [22], of the topic detection algorithm that evaluates hashtags; hence, obtaining an extended model of credibility that considers four dimensions—text, user, social, and topic credibility measures. (ii) A comparative evaluation—in terms of precision, recall, and F1 metrics—of several topic detection algorithms, based on sequential k-means, latent semantic indexing (LSI), non-negative matrix factorization (NMF), and latent Dirichlet allocation (LDA). (iii) The implementation of the extended model within the T-CREo framework [41], with the topic credibility measure based on NMF, derived from the best results of the comparative study, which allows a quantitative and qualitative evaluation of our extended credibility model in the context of detection of fake news.

This paper is structured as follows: Section 2 presents relevant works related to topic analysis and credibility. Section 3 describes the methods used for the topic analysis including topic detection algorithms, metrics for model evaluation, and distance measures. Section 4 presents the comparative evaluation of the topic detection algorithms considered. Section 5 shows the extended credibility model. Section 6 presents the qualitative and quantitative evaluation of our approach. Finally, Section 7 presents the conclusions and future work.

2. Related Work

The need for a topic detection system, particularly associated with Twitter, is motivated by the amount of information in microblogs and its use to spread information and express opinions. With this massive amount of Twitter data, users can get saturated and miss important topics [33]. Then, topic detection, as a technique for discovering the main topics automatically, can help in many applications that analyze Twitter. Therefore, in the literature, there exists a huge number of scientific papers focused on topic detection on Twitter. For example, by a simple search of the Scopus database with the keywords *topic + detection + Twitter* between 2009 to 2022, 1692 related articles are obtained. In this section, we describe only a few of the most relevant studies related to our work.

There is a large variety of techniques used for topic analysis on Twitter. Works that use machine learning techniques are basically based on supervised learning [27,29,34,42,43] or consider a hybrid approach, which includes latent Dirichlet allocation (LDA) [44] or graphs [28]. Sentiment analysis and topic detection on Twitter have been combined with the Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (a technique that simplifies the inference process assuming that each document is the result of a single topic) to analyze the content related to COVID-19 from Brazil and the USA [30]. Pattern mining (a frequent pattern mining technique), which takes frequency and utility into account at the same time, has also been used for topic detection in Twitter streams [45]. Lee et al. [43] classified Twitter Trending Topics into 18 general categories, such as sports, politics, technology, by using a text-based classification with a Bag-of-Words approach and a network-based classification. Mottaghinia et al. [34] explored different approaches to detect topics of tweets and classified these approaches into four classes of categories: with word embedding or without word embedding, specified or unspecified, offline or online, and supervised or unsupervised. The authors summarized their advantages and disadvantages and concluded that depending on the application, one of the categories may be more suitable than the other. These works are examples of the different approaches that can be used to detect topics on Twitter; however, they do not propose credibility models based on the detected topics, as we do in our proposal.

Many works focused on Twitter credibility analysis have considered topic detection as an import aspect to improve the calculation of the credibility level of tweets. In the context of the Great Eastern Japan Earthquake, Namihira et al. [27] proposed a method based on the topic and opinion classification for automatically assessing the credibility of information. For this, the authors calculated the ratio of the positive opinions to all opinions about a topic. To identify the topic of a tweet and generate a topic model, the LDA algorithm was used, and to identify if an opinion of the tweet was positive or negative, a sentiment analysis algorithm was applied. Hamdi et al. [28] proposed an approach to evaluate information sources of fake news (SOFN) in term of credibility on Twitter, based on user features (e.g., created at, name, default profile, default profile image, favorites count, statuses count, description), social graph of users (*followers/following* graph), and topic annotations. Binary Machine Learning classifier models are fed with these features to predict SOFN. A web interface framework, implemented as a web plug-in system, was proposed by Tan S. [29] to compare tweets to relevant news headlines. Considering that the news headlines are true, tweets were classified as entailment, neutral, or contradiction with respect to a specific topic. To do so, the author used four classification models: logistic regression model based on count vectorizer, support vector machine based on text content features, a feedforward network using GloVe word embeddings, and an RNN-based LSTM sentence encoder with a multilayer perceptron classifier. Yang et al. [3] designed a crowdsourcing-based credibility framework for Twitter in the context of disaster-awareness situations. This framework is able to calculate in real-time the topic-level credibility (i.e., emergency situations), by analyzing the text, linked URLs, number of retweets, and geographic information extracted from both a tweet's text and external URLs, which are kept in a database. Thus, the credibility of a detected event is increased when multiple sources (the three factors) refer to the same event: tweets, the linked URLs, and retweets referring to the same event. Thus, the tweet credibility score is calculated based on the information contained in its text and URL, and the accumulated credibility score for each event is calculated based on the number of tweets and retweets associated with the same event. Similar to these works, many other studies propose credibility models related to a specific topic. The topic does not influence the credibility level; however, it is used to classify or filter the tweets. In our work, we identify the topic of the tweet, both to filter the tweet and to impact the level of credibility. Moreover, we also analyze other aspects of the tweet to determine the level of credibility.

Some of these works have used URLs present in the tweet to support the topic analysis. Similarly, other studies consider other aspects in the tweet. Alrubaian et al. [46] proposed a hybrid approach to credibility analysis to identify implausible content on Twitter and prevent the proliferation of fake or malicious information. For this, they designed an automated classification system with four components: a reputation-based component, a credibility classifier engine, a user experience component, and a feature rank algorithm. The classifier engine component distinguishes between credible and noncredible content from a user tagged dataset considering extracted features at the tweet-, user-, and hybrid-level. These features include structural aspects of the tweet, such as length, number of tags, mentions, positive and negative words, and URLs and hashtags. The classifier used for this component is the naïve Bayes classifier with a feature rank process. Shao at al. [18] introduced Hoaxy, a platform for the collection, detection, and analysis of online misinformation and its related fact checking efforts. The platform collects and tracks misinformation and fact checking. The components of this platform consist of a monitor, a database, and different data sources (social networks and news sites). The monitor has a URL tracker for the Twitter API and a set of crawlers for both fake news and fact checking websites. The extraction of social networks is performed via a stream API and, for the news sites, use an RSS (Rich Site Summary) Parser and Scrapy Spider technologies. Then, the collected data are stored in a database for future analysis. The aim of this work was to characterize the relation between the overall social sharing activity of misinformation and fact checking.

Similar to these works, we consider hashtags as an extra factor in the topic analysis. In contrast with all the works described in this section, we propose the use of Hellinger distance for comparing the semantic proximity between the topic associated with a tweet and the topic associated with its hashtags. The topics are obtained using the NMF model, which produced better results than other topic detection algorithms such as K-means, latent semantic indexing (LSI), and latent Dirichlet allocation (LDA).

3. Topic Detection Methods

Ibrahim et al. [33] presented a survey about tools and approaches for topic detection from Twitter streams. They categorized the topic detection techniques into five categories—clustering, frequent pattern mining, exemplar-based, matrix factorization, and probabilistic—and evaluated their performance using three Twitter datasets. In terms of precision, the best results were obtained with Soft Frequent Pattern Mining and Bngram, a cluster technique; while considering recall, the best results were obtained with Column Subset Selection, a matrix factorization technique. A good balance between recall and precision was obtained with an exemplar-based topic detection model. Considering this survey, we implement and test several of the categorized algorithms to select one to be integrated into the credibility model.

We have considered three categories of techniques, which include clustering, matrix factorization, and probabilistic techniques. The algorithms were selected considering the available libraries to implement them, basically, scikit-learn (https://scikit-learn.org/stable/, accessed on 1 June 2022) and their results obtained in similar contexts. The algorithms used are sequential k-means for clustering techniques, non-negative matrix factorization (NMF) and latent semantic indexing (LSI) for matrix factorization techniques, and LDA for probabilistic techniques.

3.1. Clustering Models

Sequential k-means [47] is one of the most popular clustering techniques. This algorithm groups observations into k groups based on their characteristics. So, we can partition n observations into k clusters, $S = S_1, S_2, \dots, S_k$, such that the cluster distance (WC) is minimized [33].

$$WC = \min_{S} \sum_{i=1}^{k} \sum_{x \in Si} ||x - \mu i||^2 \tag{1}$$

Equation (1) is an iterative process that we can explain as follows:

- Initialization: Select k random points as representative centroids.
- Repeat until convergence:
  - Assign each data point to the cluster of the nearest centroid.
  - Recompute each cluster centroid as the average of the assigned points.

The sequential k-means is able to update the existing clusters by applying only new data-points [48].

3.2. Matrix Factorization Model

From these techniques, we are interested in the latent semantic indexing (LSI) and non-negative matrix factorization (NMF) algorithms.

3.2.1. Latent Semantic Indexing

LSI [49] is a popular text analysis technique. To extract the conceptual content of a document, it is necessary to establish associations between those terms that occur in similar contexts [50]. Thereby, the main idea is to match topics by concepts instead of by terms [51].

Given a data matrix X_{nxd} (n documents and d terms), LSI factorizes it to the multiplication of three matrices UDV^T, which is known as singular value decomposition (SVD) [52], as shown in Equation (2).

$$X = UDV^T \tag{2}$$

This can be interpreted as projecting the data matrix X into a lower-dimensional space whose bases are latent topics.

3.2.2. Non-Negative Matrix Factorization (NMF)

LSI has two disadvantages: (i) the factorized matrices may have negative values that do not have intuitive interpretation; (ii) the bases are latent and cannot be easily interpreted [33]. In contrast, NMF [53] is another class of techniques that guarantees that the factorized matrices contain non-negative values.

Figure 1 shows a graphical representation of this process, also represented in Equation (3) [54]. Matrix X is projected into a lower-dimensional space spanned by a set of latent topics, where the coefficients of each document with respect to these bases are contained in the rows of the matrix W and each base is represented by one row in the matrix H [33].

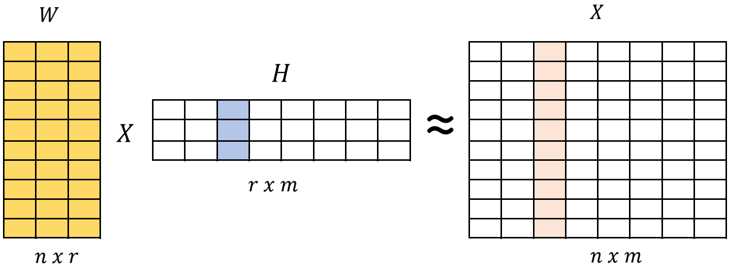


Figure 1. Non-negative matrix factorization.

$$X \approx W \times H \tag{3}$$

3.3. Probabilistic Model

We have considered LDA among the probabilistic models. LDA is a probabilistic topic modeling approach, where the document is considered as a combination of several topics and the characteristics of every topic are determined by word distribution [55]. Figure 2 shows an illustration of the definition of LDA [56], which consists of select M documents and each document contains a vector θ of topic proportions. Each word w is generated by first choosing a topic z from a multinomial parameterized by θ and then choosing a word from a multinomial conditioned on the selected topic. In our case, we treat each tweet as if it were one of the M documents.

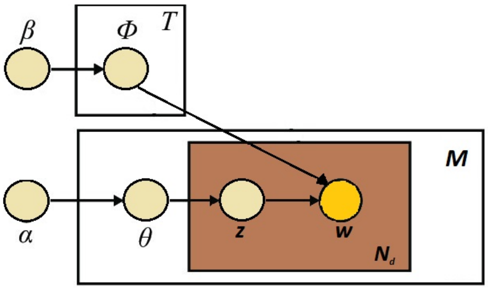


Figure 2. The latent Dirichlet allocation (LDA) model.

LDA has poor performance in the case of short texts [21] because the topics learned from this algorithm are formally a multinomial distribution over words and only the top words are used to identify the subject area or give an interpretation of a topic.

4. Comparative Evaluation of Topic Detection Algorithms

The primary goal of having the new measure associated with topic analysis is to improve the credibility model calculation using topic modeling algorithms. Our topic analysis is based on the best topic detection model considering different metrics for comparison. For this, our methodology involves preprocessing, processing, and tuning steps for each algorithm and an evaluation process to compare them. Figure 3 shows a summary of the steps of the process followed in this evaluation phase. Although we have evaluated only four topic detection algorithms, this methodology can be followed to evaluate other approaches, which we intend to conduct as future work.

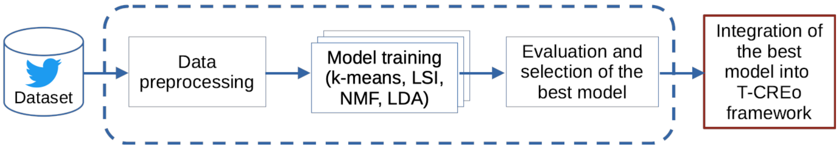

Dataset
Data preprocessing
Model training (k-means, LSI, NMF, LDA)
Evaluation and selection of the best model
Integration of the best model into T-CREo framework

Figure 3. Diagram of the comparative evaluation phase.

4.1. The Dataset Description

The dataset used for this analysis corresponds to data collected by Quezada et al. [57], a CC BY 4.0 licensed public access database (https://figshare.com/articles/dataset/tweets_csv_gz/3465974, accessed on 1 June 2022). The data contain tweets gathered from news headlines from a manually curated list of well-known news media accounts (e.g., @CNN, @BreakingNews, @BBCNews) on Twitter. The dataset is composed of a total of 43,256,261 tweets distributed across 5234 different events.

According to the privacy and tweet availability terms of Twitter (https://developer.twitter.com/en/developer-terms/agreement-and-policy), accessed on 1 June 2022, most available datasets only provide the id of tweets, as it is necessary to use the Twitter API to extract their actual texts. This API has a limit (https://developer.twitter.com/en/docs/twitter-api/rate-limits, accessed on 1 June 2022) of 900 tweet requests every 15 min; therefore, we only collected 2,000,000 tweets to create our dataset—this took us almost 1 month, managing to obtain only 86,400 tweets per day, which compose our dataset.

The obtained dataset resulted as imbalanced: in some cases, an event has less than 1000 tweets and in other cases an event has more than 100,000 tweets. In order to obtain a balanced dataset, we selected 250 events (which will be used later as topics) with 8000 tweets each. This balanced dataset is a subset from the original dataset and contains 2,000,000 tweets for training and testing purposes. For the event (topic) selection, we have taken into account that each topic to be included in our balanced dataset has to satisfy that at least 8000 tweets belong to the topic and these 8000 tweets also contain 400 tweets with at least 1 hashtag.

This dataset was structured by a tweet id and a topic id. The `tweet id` corresponds to the internal id provided by Twitter and the `topic id` is the original identifier provided by the dataset between 1 and 5234 associated to an event (or topic for us). Note that some topic identifiers are not considered in our subset since only 250 topics were selected. Table 1 shows an example of a list of the six first tweet ids in our dataset.

Table 1. Dataset structure.

Tweet id	Topic id
xxxxxx69185540096	1
xxxxxx462185543091	2
xxxxxx365534545634	2
xxxxxx435353345345	2
xxxxxx534986734857	3
xxxxxx837593759879	3

4.2. Preprocessing

The tweets have very unstructured, short texts with misspelled words, irrelevant characters, emojis, unconventional syntax, hashtags, among other elements, as well as stop words, prepositions, punctuation symbols, etc. that make more difficult the task of topic detection algorithms. For that reason, it is necessary to clean the data as part of the topic detection process. Therefore, the first step in the process is to clean the text from irrelevant words, such as usernames, URLs, emojis, and invalid characters. After that, the cleaned tweets have to be converted to a format suitable as input for the algorithms. The format that we used is the TF-IDF (Term Frequency-Inverse Document Frequency). In this preprocessing step, the following tasks are executed:

- Tokenization: the text is split at each blank character to create a list of single tokens (stand-alone words, numbers, signs, or a concatenated string such as a URL).
- Remove mentions or usernames from tweets that begin with '@' symbol and are followed by text (e.g., @jimcramer, @apple).
- Removing special characters: the characters, such as %, *, !, [,), are removed to preserve the focus on words in every tweet.
- Removing Web URLs: URLs are not considered in our topic modeling approach because they contain unspecific and hardly interpretable information.
- Removing numbers: numbers are not considered because they generally do not contain semantically viable information for our purposes.
- Removing hashtags (e.g., #AAPL, #AppleSnob), emojis, symbols, and emoticons.
- Removing frequent words and stopwords that would not provide specific semantics. These are commonly words that do not carry distinct semantic meaning, e.g., the, an, and, what.

Table 2 shows the result of applying our preprocessing step to five random tweets. After the cleaning task, we have split our dataset into training and testing sets. The proportion was 95% for training, i.e., 7600 (tweets) × 250 (topics)= 1,900,000 tweets and 5% for tests, i.e., 400 × 250 = 100,000 tweets.

Table 2. Dataset structure after cleaning and separating hashtags.

Headline Text	Hashtags	Clean Tweet
#AAPL:The 10 best Steve Jobs emails ever...htt...	[#AAPL]	best steve job email ever
RT @JPDesloges: Why AAPL Stock Had a Mini-Flas...	[#aapl]	aapl stock mini flash crash today aapl
My cat only chews @apple cords. Such an #Apple...	[#AppleSnob]	cat chew cord
I agree with @jimcramer that the #IndividualIn...	[#IndividualInvestor, #Apple, #AAPL]	agre trade extend today pullback good see
Nobody expects the Spanish Inquisition #AAPL	[#AAPL]	nobodi expect spanish inquisit

4.3. Processing

For the processing step, we train the algorithms for topic detection explained in Section 3: sequential k-means (KMEANS), latent semantic indexing (LSI), non-negative matrix factorization (NMF), and latent Dirichlet allocation (LDA). As the dataset was previously categorized by topic, we have proceeded with a supervised learning approach. The algorithms were executed iterating several times over different configurable variables to obtain the best results considering the evaluation metrics. The hyperparameters used are summarized in Table 3. The rest of the parameters corresponds to default values for each algorithm (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition/, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans/, accessed on 1 June 2022). In the case of LDA, if the data size is large, the *online update* parameter will be much faster than the *batch update* parameter. The experiments were performed using Python 3.8.10 and libraries sklearn 1.1.1, on a computer with 16 GB memory, 8 AMD vCPUs, an 80 GB disk, and SFO3-Ubuntu 20.04 (LTS) ×64.

Table 3. Hyperparameters for k-means, LSI, NMF, and LDA algorithms.

Model	Parameters	Algorithms

LDA	n_components = 250, max_iter = 100,
	learning_method = 'online'
		LatentDirichletAllocation.html

LSI	n_components = 250, n_iter = 100	TruncatedSVD.html
NMF	n_components = 250	NMF.html
KMEANS	n_clusters = 250	KMeans.html

4.4. Evaluation Metrics for the Models

We have used classical metrics to measure the models' performance, which are as follows [58]:

- Precision: The precision is the ratio $tp/(tp+fp)$, where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier to not label a negative sample as positive.
- Recall: The recall is the ratio $tp/(tp+fn)$, where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.
- F1-score: The F1 score can be interpreted as a weighted harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. F1 score is defined as $\frac{2\times precision\times recall}{precision+recall}$.

Table 4 summarizes the results of the considered algorithms, i.e., K-means, LSI, NMF, and LDA with their metrics (precision, recall, and F1 score). The evaluations were executed considering our testing set composed by 400 × 250 = 100,000 tweets. Figure 4 shows a graphical comparison between the metrics of each model generated. In terms of precision, recall, and F1, the NMF shows the best results with 0.74, 0.76, and 0.75, respectively. The second best algorithm is LDA with 0.71, followed by K-means and LSI. According to these results, the NMF algorithm is the most suitable for topic analysis in this scenario.

Table 4. Results of the algorithms: K-means, LSI, NMF, LDA.

Metrics	K-Means	LSI	NMF	LDA
Precision	0.63	0.58	0.74	0.71
Recall	0.63	0.69	0.76	0.71
F1 Score	0.63	0.63	0.75	0.71

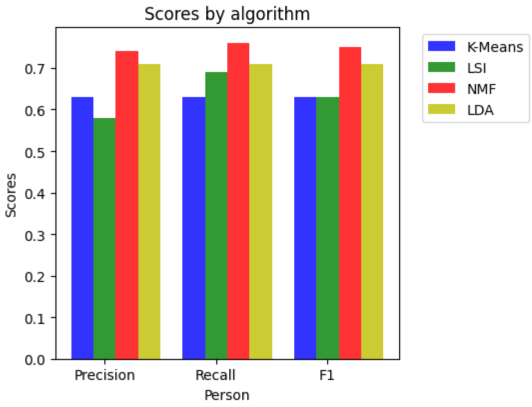


Figure 4. Results of the algorithms: K-means, LSI, NMF, LDA.

Note that for the comparative evaluation phase, the topic detection models were trained without hashtags; thus, the prediction of the topic is only based on the cleaned text. In the proposed topic detection measure, the topic of hashtags are also identified and compared with the one predicted from the plain text. The following section describes the original credibility model proposed in [22] and the extension proposed in this work by considering the topic credibility measure based on the analysis of hashtags.

5. An Extended Credibility Model Proposal: Adding Topic Measure

The credibility model proposed in [22], takes into account three measures: (i) *Text Credibility*; (ii) *User Credibility*; and (iii) *Social Credibility*. Each part has a value of 33%, by default, to compute 100% of the tweet credibility level. In the following, we first describe the original credibility model [22]; afterward, we present the extension of this model by considering topic detection.

5.1. Original Credibility Model

Figure 5 shows a general view of the original credibility model. *Text Credibility* is entirely related to the post's text, while *User Credibility* and *Social Credibility* are calculated using users' attributes. Each credibility measure is based on several components that we call filters. Hence, the model becomes easy to implement, flexible, and extensible. It does not need advanced data manipulation, which makes it ideal to use on real-time applications.

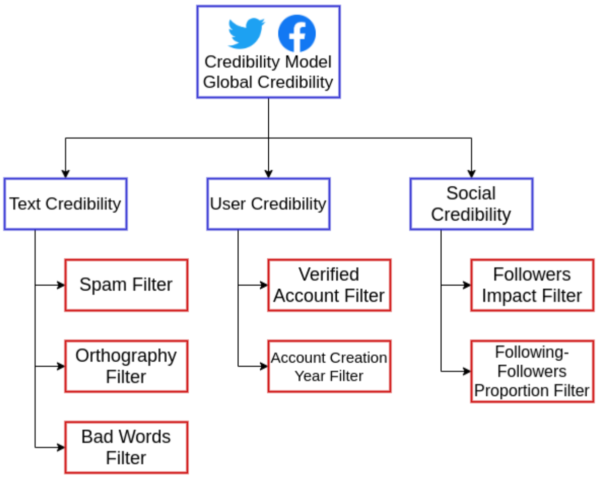


Figure 5. Original credibility model [22].

5.1.1. Text Credibility

Text credibility analyzes syntactically the content of the post (without checking the author attributes), through SPAM, bad words, and misspelling *filters*, as shown in Definition 1.

Definition 1. Text Credibility (*TextCred*). *Given the text of a post, $p.text$, Text Credibility is a function, denoted as $TextCred(p.text)$, that returns a measure $\in [0,100]$, defined as*

$$TextCred(p.text) = w_{SPAM} \times isSpam(p.text) + w_{BadWords} \times bad\_words(p.text) + w_{MisspelledWords} \times misspelling(p.text)$$

where

- *isSpam(p.text) is a SPAM detector that determines the probability $\in [0,100]$ of p.text being spam;*
- *$bad\_words(p.text)$ measures the bad words proportion $\in [0,100]$ against the number of words in a text;*
- *$misspelling(p.text)$ measures the misspelling errors proportion $\in [0,100]$;*
- *$w_{SPAM}, w_{BadWords}$, and $w_{MisspelledWords}$ represent user-defined parameters to indicate the weights that the user gives to each filter, such that $w_{SPAM} + w_{BadWords} + w_{MisspelledWords} = 1$.*

5.1.2. User Credibility

User credibility analyzes only the user as a unit of the platform, without being influenced by other users, as it is described in Definition 2.

Definition 2. User Credibility ($\boldsymbol{UserCred}$). *Given a set of metadata of a user who published a post, $p.user$, User Credibility is a function, denoted as $UserCred(p.user)$, that returns a measure $\in [0, 100]$, defined as*

$$UserCred(p.user) = Verif\_Weight(p.user) + Creation\_Weight(p.user)$$

where

- $Verif\_Weight(p.user)$ *is a function that returns* 50 *if the user is verified and* 0 *otherwise;*
- $CreationWeight(p.user)$ *measures the time since the user's account was created, with a value between* 0 *and* 50*, increasing with the longevity of the account, such as* $CreationWeight(p.user) = \frac{Account\_Age(p.user)}{Max\_Account\_Age(p.user)} \times 50$ *where*
  - $Account\_Age(p.user) = CurrentYear - YearJoined(p.user);$
  - $Max\_Account\_Age(T) = CurrentYear - SocialPlatform\_Creation\_Year;$
  - $SocialPlatform\_Creation\_Year$ *is the year in which the targeted social platform was created (e.g., 2006 for Twitter).*

5.1.3. Social Credibility

Social credibility is focused on the relations between a user account and the other accounts on the social media platform. It considers the number of *followers* and *following* (see Definition 3).

Definition 3. Social Credibility ($\boldsymbol{SocialCred}$). *Given a set of metadata of a user who published a post, $p.user$, Social Credibility is a function, denoted as $SocialCred(p.user)$, that returns a measure $\in [0, 100]$, defined as*

$$SocialCred(p.user) = FollowersImpact(p.user) + FFProportion(p.user)$$

where

- $FollowersImpact(p.user) = \frac{\min(p.user.followers, MAX\_FOLLOWERS)}{MAX\_FOLLOWERS} \times 50$ *measures the impact* $\in [0, 50]$ *on the number of followers;*
- $FFProportion(p.user_{social}) = \frac{p.user.followers}{p.user.followers + p.user.following} \times 50$ *measures the proportion* $\in [0, 50]$ *between the number of followers and followings of the user.*
- $MAX\_FOLLOWERS$ *is a user-defined parameter.*

The $MAX\_FOLLOWERS$ constant is supplied by the user, for example, in [22] it is considered as 2 million. $FFproportion$ is self-explanatory—a simple proportion that increases the credibility if the user has more *followers* than *followings*. The purpose of this function is to discredit bots, which tend to have more *followings* than *followers*.

5.1.4. Credibility Level

The credibility of a post is a weighted sum of the three credibility measures described previously. Definition 4 shows how it is calculated. According to the social network, the respective features for *User Credibility* and *Social Credibility* have to be identified and obtained.

Definition 4. Credibility Level ($\boldsymbol{Cred}$). *Given a post, p, the Credibility Level is a function, denoted as $Cred(p)$, that returns a measure $\in [0, 100]$ of its level of credibility, defined as*

$$Cred(p) = weight_{text} \times TextCred(p.text) + weight_{user} \times UserCred(p.user) + weight_{social} \times SocialCred(p.user)$$

where

- $weight_{text}$, $weigh_{user}$*, and* $weight_{social}$ *are user-defined parameters to indicate the weights that the user gives to Text Credibility, User Credibility, and Social Credibility, respectively, such that* $weight_{text} + weight_{user} + weight_{social} = 1$*; by default, they are around 33%;*
- $TextCred(p.text)$, $UserCred(p.user)$*, and* $SocialCred(p.user)$ *represent the credibility measure related to the text, the user, and the social impact of p, respectively.*

5.2. Extended Credibility Model with Topic Credibility

Once we have obtained the best trained model for topic detection, as we explain in Section 4 and Figure 3, we propose the analysis of hashtags of tweets to support topic detection. Figure 6 shows the process followed. The tweet is preprocessed and split into words to identify hashtags from the text. Then, with the topic detection model selected in the previous comparative evaluation phase, the topic of the text is determined and the topics of the hashtags are identified by applying the same process as the text, i.e., with each hashtag as an input of the NMF algorithm. Afterward, a similarity measure is calculated between the topics of the text and the topics of hashtags. We can compare if the hashtags used in a tweet are coherent with the topic treated on it. To define how *coherent* a tweet is with respect to its hashtags, we use the Hellinger distance.

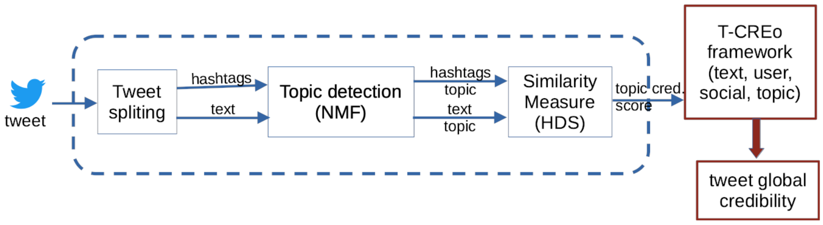


Figure 6. Topic detection process.

The Hellinger distance is a metric to measure the difference between two probability distributions. It is the probabilistic analog of Euclidean distance [59]. The Hellinger distance forms a bounded metric on the space of probability distributions over a given probability space. When comparing a pair of discrete probability distributions, the Hellinger distance is preferred because P and Q are unit length vectors as per the Hellinger scale [60]. This metric distance has been applied to other similar problems, e.g., to calculate similarity between topics [61], to find the distance between two documents [62], or to compare the distance between Tweet Corpora [59]. Due to the fact that the output of our model is a unidimensional vector with a probability distribution of topics associated with a tweet or hashtag, the Hellinger is appropriate for our problem. This distance allows calculating the semantic proximity between the topic associated to a tweet and the topic associated to its hashtags. Then, when the Hellinger distance score (HDS) approaches 1, the topics diverge, and therefore become vaguely related; when the score approaches 0, the topics become closely related.

Let us consider $f(x)$ and $g(x)$ as absolutely continuous functions. The square of the Hellinger distance is defined as shown in Equation (4) [63], where f and g are constrained to be probability density functions that integrate to 1 by definition.

$$HDS^2(f,g) = \frac{1}{2}\int (f^{1/2}(x) - g^{1/2}(x))^2 dx \tag{4}$$

Using these functions, it is possible to expand the square in the integral and obtain an alternative form for two probability distributions, P and Q, as shown in Equation (5).

$$HDS(P,Q) = \frac{1}{\sqrt{2}}\|\sqrt{P} - \sqrt{Q}\|_2 \tag{5}$$

where

- P = probability distribution for the cleaned text;
- Q = probability distribution for the hashtag.

To interpret the results of the Hellinger distance, we used the dissimilarity score (Figure 7) [64]. This means if the result is closer to 1, the dissimilarity is high; otherwise, if the result is closer to 0, the dissimilarity is low.

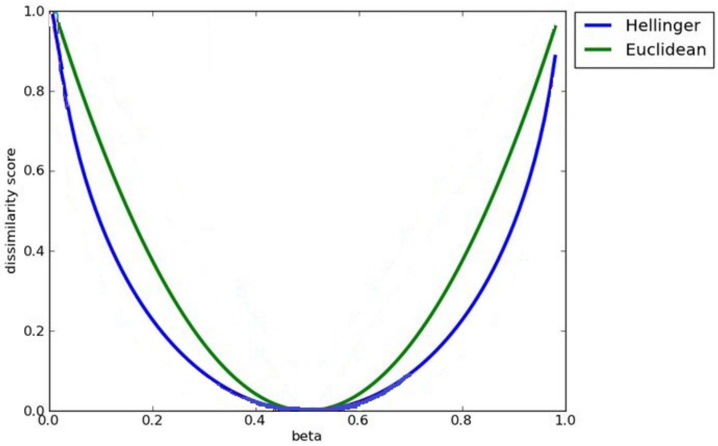


Figure 7. The behavior of the Squared Hellinger distance.

If a tweet has two or more hashtags, the model is used to obtain the topic associated to the tweet, as well as the topic associated to the individual hashtags. Afterwards, HDS will be calculated for each tweet–hashtag pair and these results will be averaged to obtain a single HDS. Definition 5 formally describes the topic credibility measure.

Definition 5. Topic Credibility (*TopicCred*). *Given the text of a post, p.text, Topic Credibility is a function, denoted as $TopicCred(p.text)$, that returns a measure $\in [0, 100]$, defined as*

$$TopicCred(p.text) = 100 \times (1 - \frac{1}{n}\sum_{i=1}^{n} HDS(NMF(p.text), NMF(p.text.hashtag_i)))$$

where

- *NMF is the topic detection algorithm;*
- *HDS is the Hellinger distance between the topics of the tweet (p.text) and the topics of the $p.text.hashtag_i$;*
- *n is the number of hashtags.*

For example, in a tweet with two hashtags, #1 and #2, the model finds the topic probability distribution of the plain text of the tweet (without the hashtags), $Ttext$, and the topic probability distribution of each hashtag, T#1 and T#2, in order to compare how far hashtag #1 and hashtag #2 are from the plain text. HDS is calculated for each hashtag, ($HDS(Ttext, T\#1)$)) and ($HDS(Ttext, T\#2)$)), and these HDS values are averaged in order to calculate the Topic Credibility measure.

Let us consider the following tweet: "*Black teenage boys are not men. They are children. Stop referring to a 17 year old as a man. You are killing children. #ferguson*". The trained NFM model is applied to the text "Black teenage boys are not men. They are children. Stop referring to a 17 year old as a man. You are killing children", as well as to the hashtag "ferguson". The result of each calculation is a vector with 250 values whose sum is 1, since the model was trained by 250 topics. Then, we have $NMF(p.text) = [Ttopic_1, Ttopic_2, \ldots, Ttopic_{250}]$ for the tweet, while for the hashtag, $NMF(p.text.hashtag_i) = [Htopic_1, Htopic_2, \ldots, Htopic_{250}]$. Once the vectors are obtained, we apply the Hellinger distance:

$$\frac{1}{\sqrt{2}} \times \|\sqrt{[Ttopic_1, Ttopic_2, \ldots, Ttopic_{250}]} - \sqrt{[Htopic_1, Htopic_2, \ldots, Htopic_{250}]}\|_2$$

The NMF("Black teenage boys are not men. They are children. Stop referring to a 17 year old as a man. You are killing children".) is $[1.25759751e^{-04}, 0.00e^{+00}, \ldots, 0.00e^{+00}]$, while for $NMF(``ferguson")$, it is $[0.00e^{+00}, 0.00e^{+00}, \ldots 2.54231396e^{-02} \ldots, 0.00e^{+00}]$; then, the HDS results to 0.3267. This similarity measure represents the topic credibility score that feeds the global credibility model in T-CREo framework (final step in Figure 6).

Our new credibility model is composed of the *Text Credibility*, *User Credibility*, *Social Credibility*, and *Topic Credibility*, as shown in Figure 8, where the Hashtag Filter represents the process of topic detection, described in Figure 6. However, there are scenarios where the trained model is incapable of assigning a topic to a given tweet or hashtag. If the percentage of association of a certain term with each of the topics allocated on the model is at or below a certain threshold (in this case, set to 0.05) the term cannot be assigned to a topic; therefore, HDS cannot be calculated. Given this scenario, the topic detection parameter should not be considered in the credibility model. By following the previous scenario, the new credibility measure is formally defined in Definition 6.

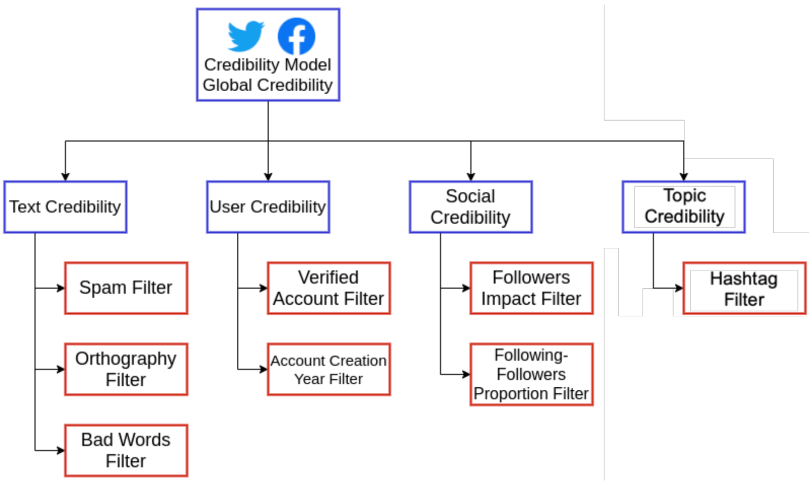


Figure 8. Extended Credibility model.

Definition 6. Credibility Level (*Cred*). *Given a post, p, Credibility Level is a function, denoted as $Cred(p)$, that returns a measure $\in [0, 100]$ of its level of credibility, defined as*

$$Cred(p) = \begin{cases} weight1_{text} \times TextCred(p.text) + \\ weight1_{user} \times UserCred(p.user) + & If\ Topic\ analysis\ is\ possible, \\ weight1_{social} \times SocialCred(p.user) + \\ weight1_{topic} \times TopicCred(p.text) \\ \\ weight2_{text} \times TextCred(p.text) + \\ weight2_{user} \times UserCred(p.user) + & Otherwise. \\ weight2_{social} \times SocialCred(p.user) \end{cases}$$

where

- *$weight1_{text}$, $weigh1_{user}$, $weight1_{social}$, and $weight1_{topic}$ are user-defined parameters to indicate the weights that the user gives to Text Credibility, User Credibility, Social Credibility, and Topic Credibility, respectively, such that $weight1_{text} + weight1_{user} + weight1_{social} + weight1_{topic} = 1$;*
- *$weight2_{text}$, $weigh2_{user}$, and $weight2_{social}$ are user-defined parameters to indicate the weights that the user gives to Text Credibility, User Credibility, and Social Credibility, respectively, such that $weight2_{text} + weight2_{user} + weight2_{social} = 1$.*

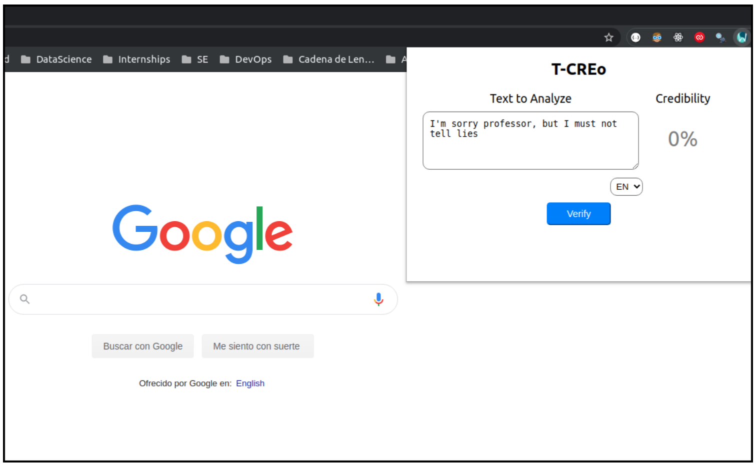
By a default configuration and under the first scenario where the topic analysis is possible, all weights—i.e., $weight1_{text}$, $weigh1_{user}$, $weight1_{social}$, and $weight1_{topic}$—are set to 25%, while for the second scenario, $weight2_{text}, weight2_{user}$, and $weight2_{social}$ are set to 33.33%.

Table 5 shows several examples of tweets that contain hashtags, the value of HDS (which shows the relation between the text and its hashtags), and the measures of credibility obtained with the original model and with the extended model. These results demonstrate that the HDS measure directly affects the credibility models if a tweet has at least one hashtag. Most of the results show that if there is a close relationship between the text and the hashtag, the credibility increases. On the other hand, if there is a far relationship between the text and the hashtag, the credibility decreases. Note that tweet #3 has an HDS value very close to 0, since Gurlitt is a composer who owns several art works. Therefore, as the text is related to the hashtag, the distance is very close (0.04) and credibility with the extended model increases (up to 71.08%), with respect to the original model (63.05%). This is unlike tweet #4, where the algorithm fails to associate the hashtag #BREAKING with the information in the text that speaks of a tragedy that occurred in Paris; therefore, credibility decreases with the extended model (70.86%) with respect to original model (74.45%). For all true tweets (#1 to #3), the extended model reports better credibility compared with the original model. For the fake tweets, the extended model decreases the global credibility in two of the three.

Table 5. Examples of HDS and credibility measures in tweets with hashtags.

N°	Tweet	Real or Fake	HDS Result	Original Model	Extended Model
1	Black teenage boys are not men. They are children. Stop referring to a 17 year old as a man. You are killing children. #ferguson	Real	0.33	55.53	58.48
2	#Putin is not the only thing missing....Look what is missing from the top of the #Kremlin today #putindead #Russia	Real	0.24	60.78	64.36
3	Tainted #Gurlitt collection should be sold with profits going to Jewish organizations. #WWII	Real	0.04	63.05	71.08
4	#BREAKING At least two killed in hostage drama east of Paris: source	Fake	0.40	74.45	70.86
5	Live Nation quashes #Prince rumour. The Purple One will not be playing at #Toronto's Massey Hall.	Fake	0.30	65.04	66.17
6	#WATCH: An aviation expert says the #4U9525 distress call was circulated on Twitter within three minutes.	Fake	0.28	74.70	73.83

The topic measure was implemented in T-CREo framework to calculate the global credibility of the tweet, which is the final step in the whole process of the topic detection, described in Figure 6. Figure 9 shows the T-CREo front-end as a Google Chrome Extension, while Figure 10 shows the credibility values under an account's timeline.



T-CREo

Text to Analyze

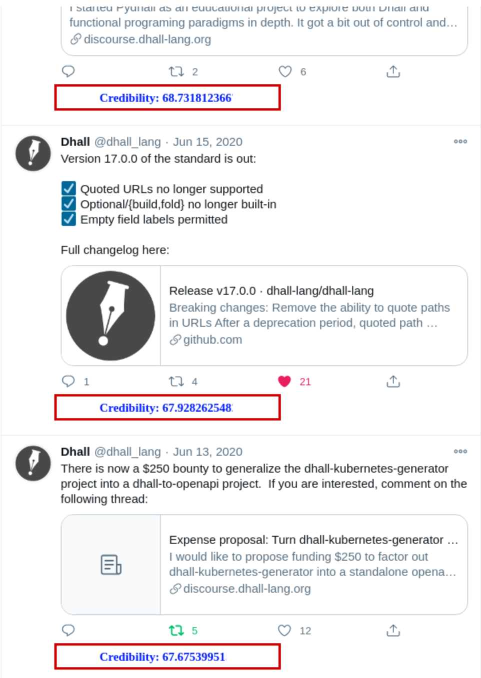
I'm sorry professor, but I must not tell lies

Credibility

0%

EN

Figure 9. T-CREo front-end as a Google Chrome Extension when opened in any website that is not Twitter.



I started Pydhall as an educational project to explore both Dhall and functional programing paradigms in depth. It got a bit out of control and…
discourse.dhall-lang.org

2 6

Credibility: 68.731812366

Dhall @dhall_lang · Jun 15, 2020
Version 17.0.0 of the standard is out:

☑ Quoted URLs no longer supported
☑ Optional/{build,fold} no longer built-in
☑ Empty field labels permitted

Full changelog here:

Release v17.0.0 · dhall-lang/dhall-lang
Breaking changes: Remove the ability to quote paths in URLs After a deprecation period, quoted path …
github.com

1 4 21

Credibility: 67.928262548

Dhall @dhall_lang · Jun 13, 2020
There is now a $250 bounty to generalize the dhall-kubernetes-generator project into a dhall-to-openapi project. If you are interested, comment on the following thread:

Expense proposal: Turn dhall-kubernetes-generator …
I would like to propose funding $250 to factor out dhall-kubernetes-generator into a standalone opena…
discourse.dhall-lang.org

5 12

Credibility: 67.67539951

Figure 10. Credibility values under an account's timeline.

The following section evaluates our model with respect to the original one.

6. Qualitative and Quantitative Evaluation

In order to evaluate our proposal, we perform a battery of experiments considering people's opinion through a survey to measure human perception and a dataset in the domain of fake news.

6.1. Qualitative Analysis

To evaluate human perception, we used the survey proposed in [65], where ten tweets were randomly selected from Twitter. This survey (the form is available at https://forms.gle/2uZNYze2YJSmCT1v7, accessed on 1 July 2022) contains opinions of 40 participants that have undergraduate and postgraduate degrees in different areas of study, using the following question *Q: How credible the following tweet is?* Then, it ranks them in a scale 1 to 10, where 1 means not believable at all and 10 means totally believable. Further, the tweets are evaluated using the original credibility model, as well as the extended version. Table 6 shows the results obtained in this test.

Table 6. Evaluation of the extended model.

Tweet ID	Survey-Avg (%)	Original Credibility Model (%)	Extended Credibility Model (%)	# Hashtags
xxxxxx9982542508038	70	68.51	68.51	00
xxxxxx6261988499456	45	44.16	44.16	00
xxxxxx454692450304	70	76.52	76.52	00
xxxxxx114923732992	15	49.74	49.74	00
xxxxxx4739103236099	65	69.78	72.33	01
xxxxxx4980596994048	30	28.99	28.99	00
xxxxxx0877124628487	50	37.67	37.67	00
xxxxxx6507817824261	45	27.86	27.86	00
xxxxxx3352350662666	65	44.05	44.05	00
xxxxxx6331631550472	40	38.69	38.69	00

The results show similar human perception values with respect to the original and extended credibility models (an average of 10.11% of difference), which validate the models. Since most of the tweets do not have hashtags, our extended credibility analysis remained almost the same as the original one, with the exception of tweet ID:xxxxxx4739103236099, which has a hashtag and where our model obtained 72.33%, while for the original model, it obtained 69.78%. To improve this evaluation, a new survey on tweets that contain hashtags is planned for future work.

6.2. Quantitative Analysis

In the domain of fake news, most of the studies apply machine learning techniques for a binary classification [66–68] (whether it is fake or not). Its evaluations are made by the use of benchmarks that consist of labeled tweets from different topics. One well-known and available dataset is PHEME, proposed by Zubia [69], which contains 6424 labeled tweets, grouped by 9 events.

Since credibility is a percentage between 0 and 100, we establish a threshold ($[0, 100]$). When the credibility value is less than the threshold, the tweet is considered as fake. To evaluate our proposal, we calculated the F1 score, based on the precision and recall, defined in Section 4.4. A variation of $step = 5$ for the threshold is used in order to evaluate several cases.

In the following sections, the results are described by event. The original model and extended model are renamed as OM and EM, respectively.

6.2.1. Event "Putin Missing"

This event has 238 tweets divided into rumors and nonrumors, of which 107 have at least one hashtag. We can see the chart of the results obtained in Figure 11. The biggest F1 score difference between the OM and the EM was obtained for the threshold of 55%, where our EM had 25.00% and the OM had 16.99%. In general, our EM had an increase of 0.17% in the F1 score. In all cases, the F1 score of the EM was equal or greater than the one of OM.

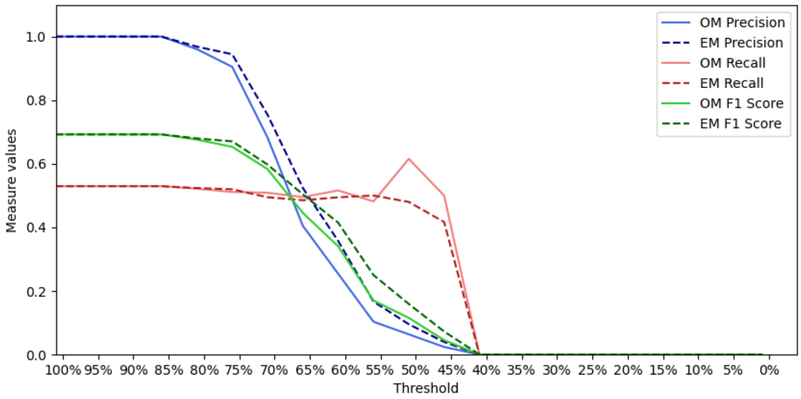


Figure 11. Results of the event "Putin missing".

6.2.2. Event "Charlie Hebdo"

This event has 2079 tweets divided into rumors and nonrumors, of which 254 have at least one hashtag. We can see the chart of the results obtained in Figure 12. For thresholds less than 40%, our EM as well as the OM resulted 0% F1 scores. The biggest difference in F1 score was for the threshold of 60%, where the OM obtained 7.45% and our proposal obtained 9.14%. In general, our model had an increase of 0.56%.

6.2.3. Event "Prince Toronto"

This event has 233 tweets divided into rumors and nonrumors, of which 83 have at least one hashtag. We can see the chart of the results obtained in Figure 13. For thresholds of 100%, 95%, 90%, and 85%, our extended model obtained the same F1 score as the original model (100%). For the threshold of 60%, we obtained the best difference (25.75% for the EM and 16.73% for the OM). In general, our EM had an increase of 1.95%.

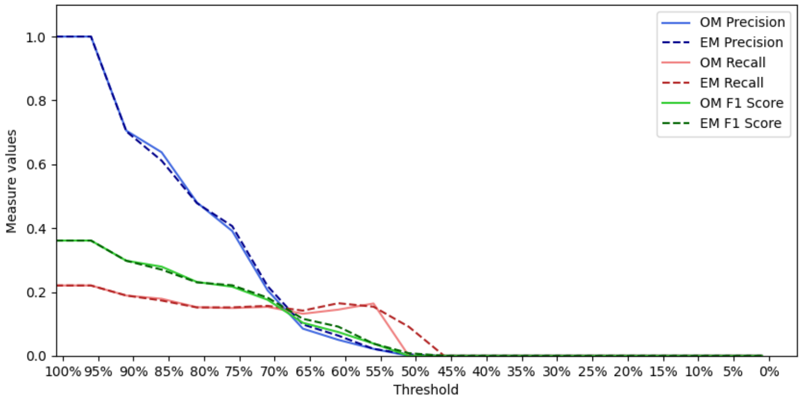


Figure 12. Results of the event “Charlie Hebdo”.

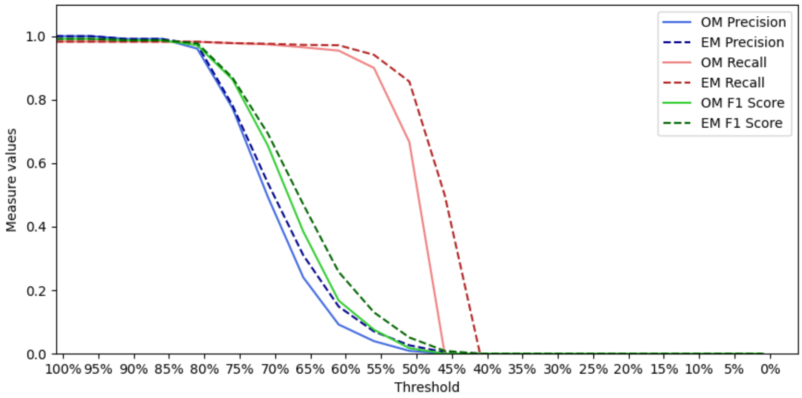


Figure 13. Results of the event “Prince Toronto”.

6.2.4. Event “Ottawa Shooting”

This event has 890 tweets divided into rumors and nonrumors, of which 221 have at least one hashtag. We can see the chart of the results obtained in Figure 14. For thresholds less than 40%, our EM and the OM obtained 0% F1 scores. The best difference was obtained for the threshold of 65%, where our EM obtained 27.63% F1 score, while the OM obtained 25.40%. In general, our model had a decrease of −0.38%.

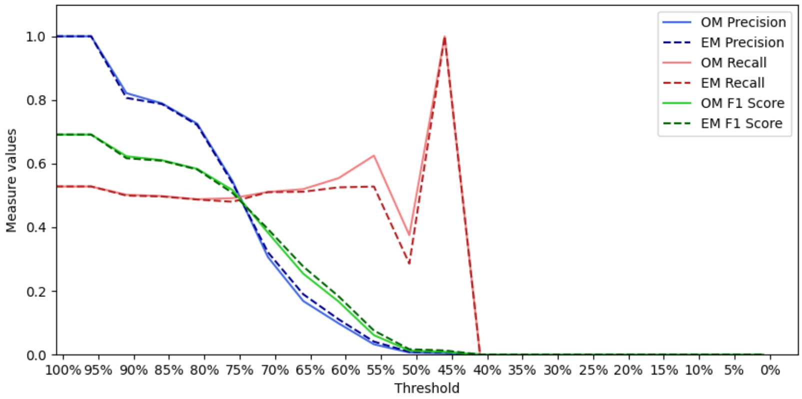


Figure 14. Results of the event “Ottawa shooting”.

6.2.5. Event "Gurlitt"

This event has 138 tweets divided into rumors and nonrumors, of which 12 have at least one hashtag. We can see the chart of the results obtained in Figure 15. For thresholds greater than 65%, our EM has a similar F1 score to the OM. The best difference was obtained for the threshold of 50%, where our EM had an F1 score of 13.69%, while the OM had 3.12%. In general, our model had an increase of 4.80%.

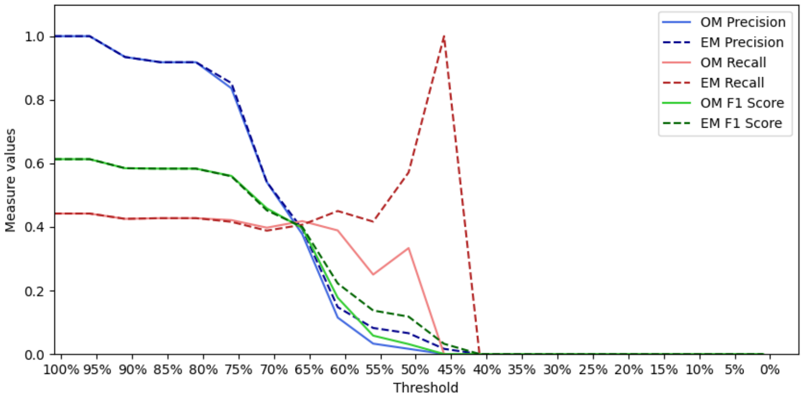


Figure 15. Results of the event "Gurlitt".

6.2.6. Event "Ebola"

This event has 14 tweets divided into rumors and nonrumors and there is no tweet that has a hashtag. Therefore, the results for both models are the same. We can see the chart of the results obtained in Figure 16.

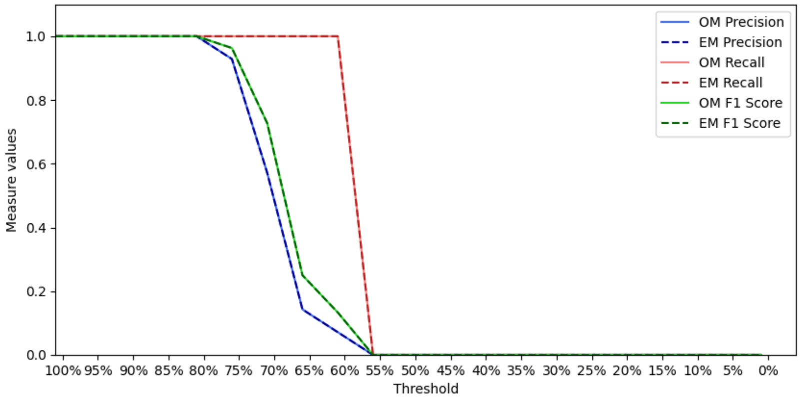


Figure 16. Results of the event "Ebola".

6.2.7. Event "Germanwings"

This event has 469 tweets divided into rumors and nonrumors, of which 30 have at least one hashtag. We can see the chart of the results obtained in Figure 17. For thresholds greater than 70%, our EM has a similar F1 score to the OM. For the threshold of 65%, our EM had the biggest difference with respect to the OM (41.92% and 40.11%, respectively). In general, our model had an increase of 0.16%.

6.2.8. Event "Ferguson"

This event has 1143 tweets divided into rumors and nonrumors, of which 1143 have at least one hashtag. We can see the chart of the results obtained in Figure 18. For thresholds less than 45%, both our EM and the OM received 0% F1 scores. The biggest difference was obtained for the threshold of 55%, where our EM had an F1 score of 11.64%, while the OM had 1.98%. In general, our model had an increase of 1.71%.

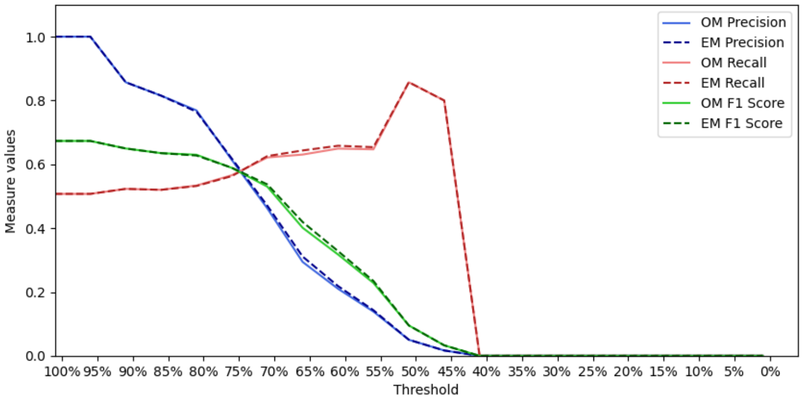


Figure 17. Results of the event "Germanwings".

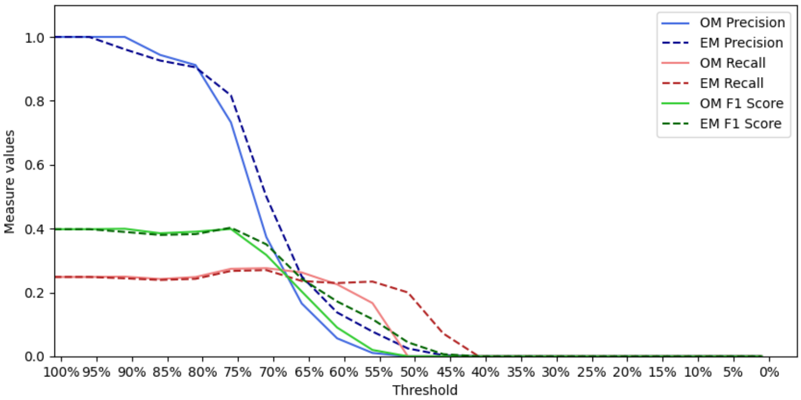


Figure 18. Results of the event "Ferguson".

6.2.9. Event "Sydney Siege"

This event has 1221 tweets divided into rumors and nonrumors, of which 137 have at least one hashtag. We can see the chart of the results obtained in Figure 19. For thresholds bigger than 80%, our EM has similar precision to the OM. The biggest difference was obtained for a threshold of 75%, where our EM had 44.44% F1 score, while the OM had 43.15%. In general, our model had an increase of 0.02%.

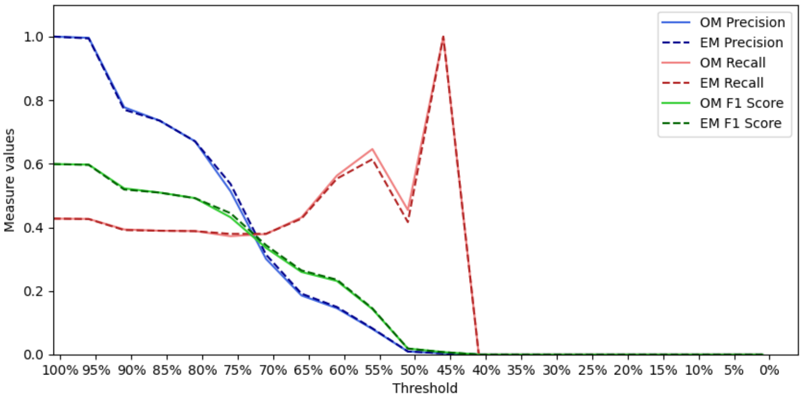


Figure 19. Results of the event "Sydney Siege".

6.2.10. All Tweets from PHEME Dataset

Figure 20 shows the precision, recall, and F1 score by thresholds of all tweets from the PHEME dataset. We can observe that our EM has better results than the OM for thresholds 45% until 75% (up to 3.04% difference for the threshold of 60%). For other thresholds, the F1 score values are similar for both models. The best F1 score was obtained with the threshold set at 95% (47.43%).

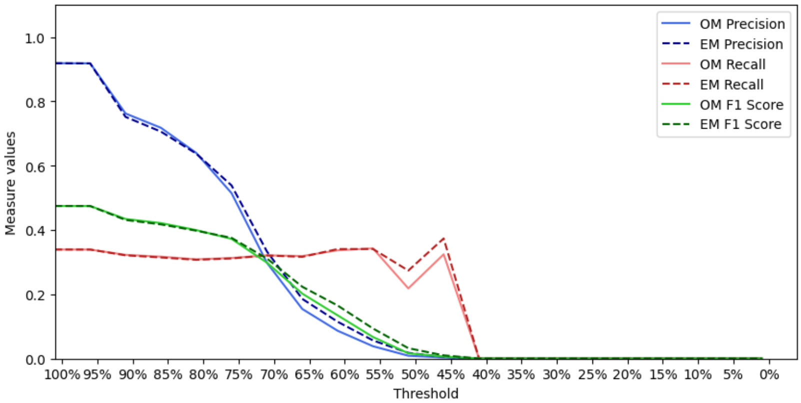


Figure 20. All Tweets from PHEME dataset by threshold.

6.2.11. All Tweets from PHEME Dataset with Hashtags

In Figure 21, we show the precision, recall, and F1 score by thresholds of all tweets that have at least one hashtag. We can observe that our EM has better results for thresholds from 40% until 75% (up to 9.60% of difference for threshold 60%). For other thresholds, the F1 score values are similar for both models. The best F1 score was obtained with the threshold set at 95% (56.41%).

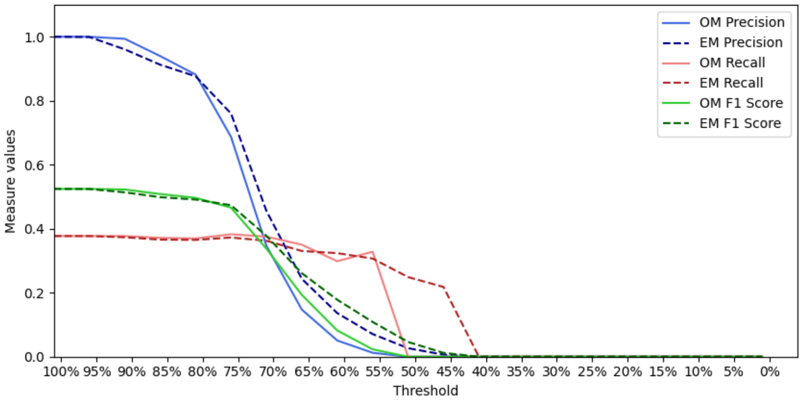


Figure 21. All tweets with hashtags from PHEME dataset by threshold.

Note that for the quantitative experiment, the initial NMF model trained by 250 topics was used, i.e., the model was not trained with the PHEME dataset; thus, better precision values can be obtained by training the model with the PHEME dataset. The idea of training the initial model with 250 topics has allowed having a generalized model that works with topics that are not included but are related to the original dataset.

7. Conclusions and Future Work

In this work, we extended a credibility model by adding topic analysis for tweets that have hashtags, since currently on Twitter it is very common to use hashtags to somehow label the tweet with words that are trending or relevant. To do so, we first compared different topic detection algorithms; we evaluated them using precision, recall, and F1 score, and we stayed with the one that gave the best results, which was NMF.

We can notice that from the 6424 tweets of PHEME dataset that only 1987 have hashtags. That means the other remaining tweets will keep their credibility probability percentage the same as the original model. It can be seen in the 'Ebola' event that since it did not contain any hashtags, the extended model obtained the same results as the original one. Moreover, the dataset with the greatest difference between the models is “Gurlitt”, due to this event being about stolen works from the Gurlitt Collection (arts collected by Cornelius Gurlitt); therefore, most of the tweets talk about art and museums, and the hashtags used for this event are also related to these words; for example: "#Entertainment", "#museum", and "#art". The extended model increases, in this event, up to 4.80% the average of F1 score values. In general, the improvement of the extended model in the PHEME dataset can be shown. Although it is not much (up to 3.04% F1 score for the 60% threshold), it is because of the fact that only 30% of the tweets had hashtags; for the case where all tweets have hashtags, the improvement is more significant (up to an F1 score of 9.60%).

Even though the NMF was not trained for PHEME dataset, we obtained high F1 score values for events such as Prince Toronto (97.80% F1 score for 80% threshold), Ebola (100% F1 score for thresholds greater than 80%), Germanwings (64.96% F1 score for 90% threshold), Ferguson (approx 39% F1 score for thresholds greater than 75%), and Sydney Siege (51.90% F1 score for 90% threshold). This effect was obtained thanks to the huge number of topics with which the model was trained, returning topics that are related.

We are currently working on extending this model by considering retweets, likes, and other attributes to measure the social impact of a tweet, which in turn could improve the measure of credibility, and applying it to other languages, such as Spanish and French. Moreover, we are planning to use a larger dataset to have a greater variety of topics and keywords that can be used for analysis, as well as to evaluate other topic detection algorithms such as neural language models and community detection.

Finally, the present study shows the feasibility of integrating a topic analysis to our credibility framework and of considering certain associated semantics. However, this concern is still a challenge. Other semantic aspects can be also incorporated such as the following: How do hashtags impact the credibility human perception of tweets? Is there coherence between the tweet image and its topic? Is the URL associated with the tweet concordant with its topic? Part of these questions will be considered as further works, with additional experiments for qualitative and quantitative assessment.

Author Contributions: Conceptualization: M.H.-M., A.A., I.D. and Y.C.; data curation: M.H.-M., A.A., I.D. and Y.C.; methodology: M.H.-M., A.A., I.D. and Y.C.; software: M.H.-M., A.A., I.D., J.C.-L. and Y.C.; investigation: M.H.-M., A.A., I.D. and Y.C.; writing—original draft preparation: M.H.-M.; writing—review and editing: M.H.-M., A.A., I.D. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the FONDO NACIONAL DE DESARROLLO CIENTÍFICO, TECNOLÓGICO Y DE INNOVACIÓN TECNOLÓGICA—FONDECYT as an executing entity of CONCYTEC under grant agreement no. 01-2019-FONDECYT-BM-INC.INV in the project RUTAS: Robots for Urban Tourism Centers, Autonomous and Semantic-based.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aksoy, M.E. A Qualitative Study on the Reasons for Social Media Addiction. *Eur. J. Educ. Res.* **2018**, *7*, 861–865. [CrossRef]
2. O'Glasser, A.Y.; Jaffe, R.C.; Brooks, M. To Tweet or Not to Tweet, That Is the Question. *Semin. Nephrol.* **2020**, *40*, 249–263. [CrossRef]
3. Yang, J.; Yu, M.; Qin, H.; Lu, M.; Yang, C. A Twitter Data Credibility Framework—Hurricane Harvey as a Use Case. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 111. [CrossRef]
4. Cooper, G.P., Jr.; Yeager, V.; Burkle, F.M., Jr.; Subbarao, I. Twitter as a Potential Disaster Risk Reduction Tool. Part III: Evaluating Variables that Promoted Regional Twitter Use for At-risk Populations During the 2013 Hattiesburg F4 Tornado. *PLoS Curr.* **2022**, *7*. [CrossRef]
5. Malik, A.; Heyman-Schrum, C.; Johri, A. Use of Twitter across educational settings: A review of the literature. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 1–22. [CrossRef]

6. Java, A.; Song, X.; Finin, T.; Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. In Proceedings of the WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD '07, San Jose, CA, USA, 12–15 August 2007; Association for Computing Machinery: New York, NY, USA, 2007. [CrossRef]
7. Antonakaki, D.; Fragopoulou, P.; Ioannidis, S. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Syst. Appl.* **2021**, *164*, 114006. [CrossRef]
8. Samuel, J.; Garvey, M.; Kashyap, R. That Message Went Viral?! Exploratory Analytics and Sentiment Analysis into the Propagation of Tweets. *arXiv* **2020**, arXiv:2004.09718. [CrossRef].
9. Walck, P. Twitter: Social Communication in the Twitter Age. *Int. J. Interact. Commun. Syst. Technol.* **2013**, *3*, 66–69.
10. Dongo, I.; Cadinale, Y.; Aguilera, A.; Martínez, F.; Quintero, Y.; Barrios, S. Web Scraping versus Twitter API: A Comparison for a Credibility Analysis. In Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services, iiWAS '20, Chiang Mai, Thailand, 30 November–2 December 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 263–273. [CrossRef]
11. Dongo, I.; Cardinale, Y.; Aguilera, A.; Martinez, F.; Quintero, Y.; Robayo, G.; Cabeza, D. A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility analysis. *Int. J. Web Inf. Syst.* **2021**, *17*, 580–606. [CrossRef]
12. Hashemi, M. The Infrastructure Behind Twitter: Scale. 2017. Available online: https://blog.twitter.com/engineering/ (accessed on 15 July 2022).
13. Markatos, E.; Balzarotti, D.; Almgren, M.; Athanasopoulos, E.; Bos, H.; Cavallaro, L.; Ioannidis, S.; Lindorfer, M.; Maggi, F.; Minchev, Z.; et al. *The Red Book*; Chalmers Research: Gothenburg, Sweden, 2013.
14. Abdullah-All-Tanvir; Mahir, E.M.; Akhter, S.; Huq, M.R. Detecting Fake News using Machine Learning and Deep Learning Algorithms. In Proceedings of the 2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 28–30 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5. [CrossRef]
15. Hassan, E.A.; Meziane, F. A Survey on Automatic Fake News Identification Techniques for Online and Socially Produced Data. In Proceedings of the 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 21–23 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6. [CrossRef]
16. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
17. Ma, J.; Gao, W.; Wong, K.F. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning. In Proceedings of the WWW '19: The World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 3049–3055. [CrossRef]
18. Shao, C.; Ciampaglia, G.L.; Flammini, A.; Menczer, F. Hoaxy: A Platform for Tracking Online Misinformation. In Proceedings of the 25th International Conference Companion on World Wide Web; International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, CHE, WWW '16 Companion, Montreal, QC, Canada, 11–15 May 2016; pp. 745–750. [CrossRef]
19. Brummette, J.; DiStaso, M.; Vafeiadis, M.; Messner, M. Read All About It: The Politicization of "Fake News" on Twitter. *J. Mass Commun. Q.* **2018**, *95*, 497–517. [CrossRef]
20. Murayama, T.; Wakamiya, S.; Aramaki, E.; Kobayashi, R. Modeling the spread of fake news on Twitter. *PLoS ONE* **2021**, *16*, e0250419. [CrossRef]
21. Mehrotra, R.; Sanner, S.; Buntine, W.; Xie, L. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013; pp. 889–892. [CrossRef]
22. Dongo, I.; Cardinale, Y.; Aguilera, A. Credibility Analysis for Available Information Sources on the Web: A Review and a Contribution. In Proceedings of the 2019 4th International Conference on System Reliability and Safety (ICSRS), Rome, Italy, 20–22 November 2019; pp. 116–125. [CrossRef]
23. Al-Khalifa, H.; Al-Eidan, R. An experimental system for measuring the credibility of news content in Twitter. *Intl. J. Web Inf. Syst.* **2011**, *7*, 130–151. [CrossRef]
24. Gupta, A.; Kumaraguru, P.; Castillo, C.; Meier, P. Tweetcred: Real-time credibility assessment of content on twitter. In Proceedings of the International Conference on Social Informatics, Barcelona, Spain, 11–13 November 2014; pp. 228–243.
25. Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; Shah, S. Real-time rumor debunking on twitter. In Proceedings of the International on Conference on Information and Knowledge Management, Melbourne, Australia, 18–23 October 2015; pp. 1867–1870.
26. AlRubaian, M.; Al-Qurishi, M.; Al-Rakhami, M.; Hassan, M.M.; Alamri, A. CredFinder: A real-time tweets credibility assessing system. In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, San Francisco, CA, USA, 18–21 August 2016; pp. 1406–1409.
27. Namihira, Y.; Segawa, N.; Ikegami, Y.; Kawai, K.; Kawabe, T.; Tsuruta, S. High Precision Credibility Analysis of Information on Twitter. In Proceedings of the 2013 International Conference on Signal-Image Technology & Internet-Based Systems, Kyoto, Japan, 2–5 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 909–915. [CrossRef]
28. Hamdi, T.; Slimi, H.; Bounhas, I.; Slimani, Y. A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding. In Proceedings of the International Conference on Distributed Computing and Internet Technology, Bhubaneswar, India, 9–12 January 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 266–280.

29. Tan, S. Spot the Lie: Detecting Untruthful Online Opinion on Twitter. Ph.D. Thesis, Department of Computing, Imperial College London, London, UK, 2017.
30. Garcia, K.; Berton, L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Appl. Soft Comput.* **2021**, *101*, 107057. [CrossRef] [PubMed]
31. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on twitter. In Proceedings of the International Conference on WWW, Hyderabad, India, 28 March–1 April 2011; pp. 675–684.
32. Lorek, K.; Suehiro-Wiciński, J.; Jankowski-Lorek, M.; Gupta, A. Automated Credibility Assessment on Twitter. *Comput. Sci.* **2015**, *16*, 157. [CrossRef]
33. Ibrahim, R.; Elbagoury, A.; Kamel, M.S.; Karray, F. Tools and approaches for topic detection from Twitter streams: Survey. *Knowl. Inf. Syst.* **2018**, *54*, 511–539. [CrossRef]
34. Mottaghinia, Z.; Feizi-Derakhshi, M.R.; Farzinvash, L.; Salehpour, P. A review of approaches for topic detection in Twitter. *J. Exp. Theor. Artif. Intell.* **2021**, *33*, 747–773. [CrossRef]
35. Alash, H.M.; Al-Sultany, G.A. Improve topic modeling algorithms based on Twitter hashtags. *J. Phys. Conf. Ser.* **2020**, *1660*, 012100. [CrossRef]
36. Huang, J.; Thornton, K.; Efthimiadis, E. Conversational Tagging in Twitter. In Proceedings of the 21st Conference on Hypertext and Hypermedia (HT), Toronto, ON, Canada, 13–16 June 2020; pp. 173–178. [CrossRef]
37. Godin, F.; Slavkovikj, V.; De Neve, W.; Schrauwen, B.; Van de Walle, R. Using topic models for Twitter hashtag recommendation. In Proceedings of the WWW '13 Companion: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 593–596. [CrossRef]
38. Kou, F.F.; Du, J.P.; Yang, C.X.; Shi, Y.S.; Cui, W.Q.; Liang, M.Y.; Geng, Y. Hashtag Recommendation Based on Multi-Features of Microblogs. *J. Comput. Sci. Tech.* **2018**, *33*, 711–726. [CrossRef]
39. Figueiredo, F.; Jorge, A. Identifying topic relevant hashtags in Twitter streams. *Inform. Sci.* **2019**, *505*, 65–83. [CrossRef]
40. Ma, Z.; Dou, W.; Wang, X.; Akella, S. Tag-Latent Dirichlet Allocation: Understanding Hashtags and Their Relationships. In Proceedings of the WI-IAT '13: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)—Volume 01, Melbourne, Australia, 14–17 December 2020; IEEE Computer Society: Piscataway, NJ, USA, 2013; pp. 260–267. [CrossRef]
41. Cardinale, Y.; Dongo, I.; Robayo, G.; Cabeza, D.; Aguilera, A.; Medina, S. T-CREo: A Twitter Credibility Analysis Framework. *IEEE Access* **2021**, *9*, 32498–32516. [CrossRef]
42. Verasakulvong, E.; Vateekul, P.; Piyatumrong, A.; Sangkeettrakarn, C. Online Emerging Topic Detection on Twitter Using Random Forest with Stock Indicator Features. In Proceedings of the 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), Nakhonpathom, Thailand, 11–13 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6. [CrossRef]
43. Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M.d.M.A.; Agrawal, A.; Choudhary, A. Twitter Trending Topic Classification. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 251–258. [CrossRef]
44. Zhang, C.; Lu, S.; Zhang, C.; Xiao, X.; Wang, Q.; Chen, G. A Novel Hot Topic Detection Framework With Integration of Image and Short Text Information From Twitter. *IEEE Access* **2018**, *7*, 9225–9231. [CrossRef]
45. Choi, H.J.; Park, C.H. Emerging topic detection in twitter stream based on high utility pattern mining. *Expert Syst. Appl.* **2019**, *115*, 27–36. [CrossRef]
46. Alrubaian, M.; Al-Qurishi, M.; Hassan, M.; Alamri, A. A Credibility Analysis System for Assessing Information on Twitter. *IEEE Trans. Dependable Secur. Comput.* **2016**, *15*, 661–674. [CrossRef]
47. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1988. [CrossRef]
48. Tehrani, A.F.; Ahrens, D. Modified sequential k-means clustering by utilizing response: A case study for fashion products. *Expert Syst.* **2017**, *34*, e12226. [CrossRef]
49. Landauer, T.K.; Foltz, P.W.; Laham, D. An introduction to latent semantic analysis. *Discourse Process.* **1998**, *25*, 259–284. [CrossRef]
50. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
51. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the SIGIR '99: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; Association for Computing Machinery: New York, NY, USA, 1999; pp. 50–57. [CrossRef]
52. Golub, G.H.; Reinsch, C. Singular value decomposition and least squares solutions. *Numer. Math.* **1970**, *14*, 403–420. [CrossRef]
53. Wang, Y.X.; Zhang, Y.J. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Trans. Knowl. Data Eng.* **2012**, *25*, 1336–1353. [CrossRef]
54. Saxena, A.; Mueller, C. Intelligent Intrusion Detection in Computer Networks using Swarm Intelligence. *Int. J. Comput. Appl.* **2018**, *179*, 1–9. [CrossRef]
55. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [CrossRef]
56. Gimpel, K. *Modeling Topics*; Technical Report; Carnegie Melon University: Pittsburgh, PA, USA, 2006.
57. Kalyanam, J.; Quezada, M.; Poblete, B.; Lanckriet, G. Prediction and Characterization of High-Activity Events in Social Media Triggered by Real-World News. *PLoS ONE* **2016**, *11*, e0166694. [CrossRef] [PubMed]

58. Godbole, S.; Sarawagi, S. Discriminative Methods for Multi-labeled Classification. In *Advances in Knowledge Discovery and Data Mining*; Springer: Berlin, Germany, 2004; pp. 22–30. [CrossRef]
59. Xin, E.Z.; Murthy, D.; Lakuduva, N.S.; Stephens, K.K. Assessing the Stability of Tweet Corpora for Hurricane Events Over Time: A Mixed Methods Approach. In Proceedings of the SMSociety '19: 10th International Conference on Social Media and Society, Toronto, ON, Canada, 19–21 July 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 59–66. [CrossRef]
60. González-Castro, V.; Alaiz-Rodríguez, R.; Alegre, E. Class distribution estimation based on the Hellinger distance. *Inform. Sci.* **2013**, *218*, 146–164. [CrossRef]
61. Maiya, A.S.; Rolfe, R.M. Topic Similarity Networks: Visual Analytics for Large Document Sets. *arXiv* **2014**, arXiv:1409.7591. [CrossRef].
62. Dingemans, S. Application of Short Text Topic Modelling Techniques to Greta Thunberg Discussion on Twitter. Master's Thesis, National College of Ireland, Dublin, Ireland, 2020.
63. Hawking, S. *Hellinger Distance—Encyclopedia of Mathematics*; EMS: Berlin, Germany, 1988.
64. Brandmaier, A. *Permutation Distribution Clustering and Structural Equation Model Trees*; Technical Report, Science and Technology Faculties; University of Saarland: Berlin, Germany, 2011.
65. Lupa, J.C. Análisis de Credibilidad en la Red Social Twitter a través de su Actividad Social. Bachelor's Thesis, Universidad Católica San Pablo, Arequipa, Peru, 2021.
66. Mandical, R.R.; Mamatha, N.; Shivakumar, N.; Monica, R.; Krishna, A.N. Identification of Fake News Using Machine Learning. In Proceedings of the 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2–4 July 2020; pp. 1–6. [CrossRef]
67. Aphiwongsophon, S.; Chongstitvatana, P. Detecting Fake News with Machine Learning Method. In Proceedings of the 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Rai, Thailand, 18–21 July 2018; pp. 528–531. [CrossRef]
68. Ahmed, A.A.A.; Aljabouh, A.; Donepudi, P.K.; Choi, M.S. Detecting Fake News Using Machine Learning: A Systematic Literature Review. *arXiv* **2021**, arXiv:2102.04458.
69. Zubiaga, A.; Zubiaga, A.; Hoi, G.W.S.; Liakata, M.; Procter, R. PHEME dataset of rumours and non-rumours. *Figshare* **2016**. [CrossRef]