A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility analysis

Article *in* International Journal of Web Information Systems · August 2021

DOI: 10.1108/IJWIS-03-2021-0037

CITATION
1

READS
92

7 authors, including:

Irvin Dongo
Ecole Supérieure des Technologies Industrielles Avancées (ESTIA)
21 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Yudith Cardinale
Simon Bolívar University
72 PUBLICATIONS **192** CITATIONS
SEE PROFILE

Ana Isabel Aguilera
Universidad de Valparaíso (Chile)
42 PUBLICATIONS **71** CITATIONS
SEE PROFILE

Yuni Quintero
Simon Bolívar University
2 PUBLICATIONS **4** CITATIONS
SEE PROFILE

A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility analysis

Irvin Dongo
Electrical and Electronics Engineering Department, Universidad Católica San Pablo, Arequipa, Peru and Univ. Bordeaux, ESTIA Institute of Technology, Bidart, France

Yudith Cardinale
Electrical and Electronics Engineering Department, Universidad Católica San Pablo, Arequipa, Peru and Dpto. de Computación y T.I. Universidad Simón Bolívar Biblioteca, Caracas, Venezuela

Ana Aguilera
Escuela de Ingenieria Informatica, Universidad de Valparaiso, Valparaiso, Chile, and

Fabiola Martinez, Yuni Quintero, German Robayo and David Cabeza
Universidad Simon Bolivar, Caracas, Venezuela

Received 27 March 2021
Revised 2 June 2021
Accepted 3 June 2021

Abstract

Purpose – This paper aims to perform an exhaustive revision of relevant and recent related studies, which reveals that both extraction methods are currently used to analyze credibility on Twitter. Thus, there is clear evidence of the need of having different options to extract different data for this purpose. Nevertheless, none of these studies perform a comparative evaluation of both extraction techniques. Moreover, the authors extend a previous comparison, which uses a recent developed framework that offers both alternates of data extraction and implements a previously proposed credibility model, by adding a qualitative evaluation and a Twitter-Application Programming Interface (API) performance analysis from different locations.

Design/methodology/approach – As one of the most popular social platforms, Twitter has been the focus of recent research aimed at analyzing the credibility of the shared information. To do so, several proposals use either Twitter API or Web scraping to extract the data to perform the analysis. Qualitative and quantitative evaluations are performed to discover the advantages and disadvantages of both extraction methods.

Findings – The study demonstrates the differences in terms of accuracy and efficiency of both extraction methods and gives relevance to much more problems related to this area to pursue true transparency and legitimacy of information on the Web.

Originality/value – Results report that some Twitter attributes cannot be retrieved by Web scraping. Both methods produce identical credibility values when a robust normalization process is applied to the text (i.e. tweet). Moreover, concerning the time performance, Web scraping is faster than Twitter API and it is more flexible in terms of obtaining data; however, Web scraping is very sensitive to website changes.

This research was supported by the FONDO NACIONAL DE DESARROLLO CIENTÍFICO, TECNOLÓGICO Y DE INNOVACIÓN TECNOLÓGICA – FONDECYT as executing entity of CONCYTEC under grant agreement no. 01–2019-FONDECYT-BM-INC.INV in the project RUTAS: Robots para centros Urbanos Turísticos Autónomos y basados en Semántica.

International Journal of Web Information Systems

1744-0084
DOI 10.1108/IJWIS-03-2021-0037

Additionally, the response time of the Twitter API is proportional to the distance from the central server at San Francisco.

Keywords API, Web scraping, Twitter, Credibility, Qualitative analysis

Paper type Research paper

1. Introduction

Social network platforms, such as Twitter, Facebook and Instagram have considerably increased their number of users in past years. These platforms share contents, opinions, news and sometimes fake content. In particular, Twitter is a worldwide social network which has more than 600 million users and it is one of the most widely used platforms during relevant events (Gupta *et al.*, 2014a), such as natural disasters (Gupta *et al.*, 2014b), brands and products advertising (Sánchez-Rada and Iglesias, 2019) and presidential elections (Bovet and Makse, 2019). However, the information shared on Twitter and in other social networks is not completely reliable (Zannettou *et al.*, 2019). The information posted on social networks must be reliable as their content can help people during crisis situations, influence the crowds and even can be useful as a means to help companies in decision-making. Thus, there exist many studies focused on analyzing the credibility of the shared information (Gupta *et al.*, 2014a; Alrubaian *et al.*, 2016a, 2016b; Dongo *et al.*, 2019).

In social networks, the credibility study is affected by different factors (Dongo *et al.*, 2019), such as:

- the veracity of the text, content with misspellings and bad words;
- users on the network who generated content;
- the quantity of data related to the information to be validated.

As more data are available, more features can be extracted and thus, a better analysis can be performed. In the state-of-the-art, two main extraction methods have been used:

(1) Web scraping, which consists of parsing the website HyperText Markup Language (HTML) to obtain data by using the tags; and

(2) API, which is an interface provided by social media platforms to retrieve specific information.

In previous work, we presented a framework that implements both extraction methods to perform credibility analysis (Dongo *et al.*, 2020), according to a previously presented credibility model (Dongo *et al.*, 2019). In this work, we perform an exhaustive revision of relevant and recent related studies and compare them in terms of the elements considered in the credibility measure such as text, user, social and topic and the extraction method used to get these elements (i.e. Web scraping or API). This comparison reveals that both extraction methods are currently used to analyze credibility on Twitter. Thus, there is clear evidence of the need of having different options to extract different data for this purpose. Nevertheless, none of these studies perform a comparative evaluation of both extraction techniques, as we do in this work. Moreover, we extend the previous comparison by adding a qualitative evaluation and a Twitter-API request performance from different locations around the world.

To compare both data extraction techniques, we use a similar scenario presented in Dongo *et al.* (2020) for the quantitative comparison: three different languages (Spanish, English and French) are used to evaluate their impact on the text credibility. Moreover, different types of text such as short, long, use of bad words, misspelling and emoticons are

analyzed. Social credibility is also evaluated by using two types of accounts: common accounts (less than 1,000 of followers) and famous accounts (more than 900,000 of followers). Moreover, the response time of Twitter API from different locations is evaluated. Additionally, the execution time to extract the features is also analyzed. For the qualitative comparison, several Twitter attributes and their availability using Web scraping or Twitter API are reported.

Experiments show that attributes and credibility measures retrieved by Web scraping methods are less than the ones from Twitter API. Also, a robust normalization process on the text obtained by the extraction methods produces identical credibility results. The language has no impact on the credibility, nor does the type of text. Moreover, the number of *followers* obtained by the extraction methods have a minor difference for famous accounts, as the number of *followers* is constantly growing in real-time.

Additionally, data extraction with Web scraping is faster than with Twitter API, as for the former, only local extraction (in the Twitter website) is needed, while for the API, a local extraction to obtain the user_id and tweet_id is required to latter perform a remote API request.

This paper is organized as follows. First, in Section 2, the topics related to this work such as data extraction methods and the social network Twitter, are described. Data extraction techniques are explained in Section 3. The credibility model used in this work is explained in Section 4. In Section 6, we present an experimental evaluation and a discussion about both extraction methods. Finally, we conclude in Section 7.

2. Social networks as information sources: preliminary

Social networks are online platforms that allow the free exchange of information between the users that make up their community. As the creation of social networks, they have become a new channel for communication and socialization (Alrubaian *et al.*, 2019). The types of information that are exchanged on social media platforms range from personal contexts, with the exchange of personal conversations through messaging, expressions of feelings, photographs, videos, to official and institutional contexts. Hence, social networks have become popular sources of real-time information. In 2020, among the most popular social networks in terms of the number of users are Facebook with 2,449 million, YouTube with 2,000 million, Instagram with 2,000 million, Twitter with 340 million and other large Chinese social networks, such as QQ/Qzone, Sina Weibo and Baidu Teiba, that together accumulate more than 1,000 million followers [1]. Additionally, more than 4.5 billion people now use the internet, while social media users have passed the 3.8 billion mark (Dig, 2020).

Social networks have a considerable influence on the formation of opinions due to their expansion and dissemination capacity. However, as well as they circulate correct content, they can be a way to transmit fake news and rumors. For this reason, the misuse of these platforms for harmful activities and the dissemination of disinformation is a relevant issue.

Due to the importance of the information, which cannot be underestimated at the same time it is generated, many researchers have studied the credibility on social networks, mainly based on metadata provided by themselves (Alrubaian *et al.*, 2019; Zannettou *et al.*, 2019). The techniques for data extraction include different methods or a combination of them. The next section describes these data extraction methods.

2.1 Data extraction methods

Nowadays, data extraction from Web sources is a vital task for most of the business process, research studies and others. The process of data extraction consists of obtaining relevant data or metadata useful for diverse purposes. Three well-known methods have been applied

for this: Web scraping, APIs and manual extraction. Web scraping and APIs are automated techniques and the most practical ways of data harvesting (Slamet *et al.*, 2018). They allow to collect data from various website pages and repositories, at a high speed and accurately. The data is then saved and stored for further use and analysis. Manual extraction is more susceptible to human errors and time-consuming.

2.1.1 Web scraping, use and limitations. Web data scraping can be defined as the process of extracting and combining contents of interest from the Web in a systematic way. In such a process, a software agent, a Web robot or a script, mimics the browsing interaction between the Web servers and the human in a conventional Web traversal. Step by step, the robot/script accesses as many websites as needed, parses their contents to find and extract data of interest and structures those contents as desired (Glez-Peña *et al.*, 2013). Web scrapers are useful when retrieving and processing large amounts of data quickly from a specific website. Thus, if the information is displayed on a browser, it can be accessible via a robot/script to extract the data and store them in a database for future analysis and use (Mitchell, 2015).

Web scraping is used commonly on website pages that use markup languages such as HTML or eXtensible HyperText Markup Language (XHTML). In this case, scraping consists on parsing hypertext tags and retrieving plain text information embedded onto them. The Web data scraper establishes communication with the target website through the Hypertext Transfer Protocol (HTTP) protocol and extracts the contents of interest. Some regular expression matching could be necessary along with additional logic (Glez-Peña *et al.*, 2013). The network speed may be a limitation or disadvantage for Web scraping, as it affects when and how the data is displayed. Another problem regarding this method is the often changes of the Web page format. Web scraping involves site-specific programming and does not comply with expectable changes in the HTML source (Glez-Peña *et al.*, 2013). When this happens, the whole script that gives life to the scraper must be updated and adapted to the new format or layout of the information. Web scraping maintenance is critical and could involve time-consuming programming tasks. Web scraping developers should take in consideration legal and policy issues about the information they are extracting to prevent copyright infringement (Glez-Peña *et al.*, 2013). There is no difference between visiting or scraping a website, in both cases, the user is a guest and thus, he or she is not the owner of the data extracted.

2.1.2 API as a tool, use and limitations. An API is a component of object-oriented programming languages that allow developers to build software for a particular application through a reference program library (Boillot, 2012). The API is prescribed by a device's operating system or an application program in which a requester (another device or a client user) can make requests expecting responses from them. APIs facilitate interaction between different software programs and access to their services. It includes the specification of data structures, protocols, object classes and runtimes to communicate the consumer with the resources offered by the API (Salt and Sellhorn, 2014). Developers can build new classes or extend existing ones to add new features or functionalities. A client API is called through an *endpoint*, which is a component that listens when a request is being made from the client-side of the communication to the server-side via HTTP, expecting a response to be returned.

Concerning social networks and Web information sources, many of them do not offer an API to access the available information, for several reasons (Mitchell, 2015), such as:

- the data that is wanted is small or uncommon;
- the source does not have the infrastructure or technical ability to create an API;
- the data is valuable or protected and not intended to be spread widely; and

- even when an API does exist, there may be request volume and rate limits; also, the types and format of the data that it provides might be insufficient for the purpose.

Furthermore, there are limitations on the API that include the rejection of access, if the use of the information is not enough or properly demonstrated. Data protection laws related to privacy and most of the TOS (Terms of Service) limit also their access (Dongo *et al.*, 2019).

2.1.3 Manual extraction. Other extraction method includes the manual or human extraction of features for credibility assessment. It consists on the perception of the credibility of users based on visible characteristics of posts. The study presented in Edgerly and Vraga (2019), describes an experimental design to test whether the Twitter verification mark contributes to perceptions of information and account credibility, among organizations of news. Authors show also how to account ambiguity and account congruence with political beliefs determine this relationship. Results of this study suggest that little attention is paid to the verification mark when judging credibility, even when little other information is provided about the account or the content. Instead, account ambiguity and congruence dominate credibility assessments of news organizations. Other study in this sense is presented in Vaidya *et al.* (2019), which investigates if the account verification affects the believe content of tweets. Authors found that – in the context of unfamiliar accounts – most users can effectively distinguish between authenticity and credibility. The presence or absence of an authenticity indicator has no significant effect on willingness to share a tweet or take action based on its contents.

The following section describes Twitter as an information source, which is used in this study to analyze the tweet credibility.

2.2 Twitter

Twitter has demonstrated to be one of the most populated microblogging sites in the world, with hundreds of millions of users as an excellent spread information platform in real-time (Alrubaian *et al.*, 2019). Twitter is a social network in which users share text stories, with a maximum of 280 characters and they can comment or retweet (i.e. by sharing the publication of another user).

Twitter defines itself as "what is happening in the world and the issues that people are talking about." It is currently one of the most used social networks by the media due to its flexibility of use for both users and researchers. Twitter is defined as a news media channel as much as a social platform as the relation between users does not have to be bi-directional (Alrubaian *et al.*, 2019). This platform is self-publishing based on the immediacy of users' messages. Thus, the information posted on Twitter can be spread quickly in contrast to other social networks.

Twitter API provides three types of products to extract data from tweets and accounts: Standard, Premium and Enterprise. Each one has a variety of endpoints to extract the information posted on Twitter through requests. Table 1 shows certain limitations and differences among the products offered by Twitter. Depending on the developer or project needs, the information provided by Standard API may be sufficient, but the negative point of this product is the access only to the past 7 days. For companies and important research projects, Twitter offers other products that can be adapted to customer needs such as Premium, where a subscription of $99, has 100 requests per month, 500 tweets per request and full history access, while for $149, 500 requests per month, 500 tweets per request and past 30 days of access are provided. A similar scenario can be observed for Enterprise products, which can be customized according to the requirements of the users.

Product	Price per month	Level of usage	Tweets per request	Frequency	Time frame
Standard	Free	–	Depends on endpoints	15 request per 15 mi	Past 7 days
Premium	From $149	From 500 request/ month	500	10 RPS, 60 RPM	Past 30 days
	From $99	From 100 request/ month	500	10 RPS, 30 RPM	Full history
Enterprise	Customized with predictable pricing	–	Each customer has a defined rate limit for their endpoints	APIs allow you to retrieve up to 500 results per response for a given timeframe	Past 30 days
	Customized with predictable pricing	–	Each customer has a defined rate limit for their endpoints	Default 120 RPM	Full history

Table 1.
Twitter products

Not all APIs offered by social networks are feasible for the extraction of their data. However, it has been demonstrated that Twitter API is simple and easy to use; only knowing the tweet ID or user ID, it is possible to obtain a variety of information. Twitter has different products to get their data and its structure is less complex than other social networks (Gupta *et al.*, 2018). Twitter provides only one profile and the user could choose to be private or public.

3. Related work

In this section, we describe some studies that have based their credibility analysis in Twitter either on Web scraping or the API. More related studies to our work are the ones that compare both extraction techniques in any context on social networks. We found a few of them, which are presented at the end of this section.

3.1 Data extraction methods in platforms for credibility analysis

In a previous study presented in Dongo *et al.* (2019), we describe a generic framework to calculate the credibility of several social networks. The credibility model proposed in that work is based on attributes of the post and the user account to measure three levels of credibility: text, user and social impact. We present implementation to analyze credibility in Twitter in real-time, as a Google Chrome extension application. The analysis of the texts (i.e. tweets) is further done through filters that detect spam, bad words and misspelling. Attributes such as *followers*, *following*, the joined year and the *verified account*, are considered to evaluate user and social credibility. Due to Twitter API limitations, Web scraping was used to extract from the Web pages of Twitter all these attributes. Afterward, we extended that implementation, by updating the scraper and including the use of the Twitter API to perform the comparative evaluation analysis of both extraction methods (Dongo *et al.*, 2020). In this paper, we use that implementation to perform a more exhaustive quantitative and qualitative comparison.

Hoaxy (Shao *et al.*, 2016) is a Web platform for tracking social news sharing. The system collects data from two main sources: news websites and social media, by using different technologies (i.e. Web scraping, Web syndication and, where available, APIs of social networking platforms) to populate a data set over the course of several months. The extracted data is focused on Twitter content, considering URLs and the social aspects to track the activity of the user and the tweet. This data set is analyzed considering the

temporal relation between the spread of misinformation and fact-checking and the differences in how users share them.

A systematic methodology is presented in Liu *et al.* (2015), aimed at mining language features such as people's opinion, find witness accounts, derive underlying belief from messages, use sourcing, network propagation, credibility and other user and meta-features to debunk rumors. Due to Twitter API does not let them track the propagation of retweets in as much detail, the author uses a Web scraper to get the full history and to download all tweets automatically. Their system continues monitoring the rumored event and generates dynamic real-time updates based on any additional information received.

A Web interface framework implemented as a Web plug-in system is proposed in Tan (2017). The aim is to analyze, in real-time, the credibility of tweets regarding to a specific topic. Only the text of each tweet is analyzed to be classified as being "entailment," "neutral" or "contradiction" with respect to the topic. The system shows a list of news information related to the topic; thus, users can decide the veracity of the tweet in question. The Twitter API is used to collect tweets, Web scraping is used to get the URLs referenced in tweets and the Bing news API is used to find articles and retrieve news headlines related to the topic.

A real-time system to calculate the credibility of tweets was developed in Gupta *et al.* (2014a). Authors collect data from Twitter streaming API for a set of predefined keywords in the context of six crisis events of the world during 2013. A tweet downloaded from Twitter API contains a series of fields in addition to the text, such as posting date and *followers/following* of the user at the time of the tweet. This work made a total of 1'300,000 API requests. Another real-time work is CredFinder (Alrubaian *et al.*, 2016a), which consists of a front-end in the form of a Chrome extension and a Web-based back-end. The former collects tweets in real-time from a Twitter search or a user-timeline page and the latter analyzes the collected tweets, assesses their credibility and computes a credibility score. Using the Twitter streaming API, tweets and their metadata are obtained. In the same context of real-time applications, a framework for credibility analysis is described in Iftene *et al.* (2020). This framework considers text and user credibility, aimed to identify fake users and fake news, based on neural network models. The Twitter API is used to retrieve information regarding retweets, favorites and the date of post. Also, the text is analyzed to count the number of words, number of characters, identify stop-words, etc.

A credibility analysis system for assessing information credibility on Twitter to prevent the proliferation of fake information is proposed in Alrubaian *et al.* (2016b). This work measures and characterizes the content and sources of tweets. For that, the authors designed an automated classification system with four components:

(1) a reputation-based component;

(2) a credibility classifier engine;

(3) a user experience component; and

(4) a feature rank algorithm.

They use the Twitter API to extract the features used by the system.

Considering the topic feature, a technique to enhance the ability of social network users to identify relevant sources of information (relevant, expert and useful users to follow) for a given topic is proposed in Canini *et al.* (2010). This work generates a ranked list of relevant users in combining a basic text search with an analysis of the social structure of the network. The Twitter API is used to execute this search. In Castillo *et al.* (2011), the authors propose an automatic method for assessing the credibility of a set of tweets related to "trending" topics. The tweets were collected using Twitter Monitor [2]. The data set was

manually tagged as credible or not credible. After, a supervised classifier was trained to predict credibility levels on Twitter events. Also, analyzing the topic feature, TwitterBOT (Lorek *et al.*, 2015) assesses the credibility of tweets. The data collecting is based on gathering real tweets posted on Twitter on one particular subject (nature environment preservation). A manual tagging is done on each tweet in the data set as "highly credible," "highly not credible," "neutral" and "controversial." A supervised learning model is developed using a RandomForest classifier to execute an automated credibility assessment.

Another work-related to topics is the one proposed in Namihira *et al.* (2013). Credibility is assessed by calculating the ratio of the positive opinions to all opinions about a topic. Sentiment analysis is performed using a semantic orientation dictionary to identify the opinions as positive and negative. This work considers too user's knowledge (expertise) in the information credibility assessment. Authors do not declare that Twitter API was used to collect the tweets, but their system architecture shows a collector module from Twitter. In Yang *et al.* (2019a), a framework for credibility analysis on Twitter data, with disaster situation awareness is proposed. This framework is able to calculate topic-level credibility (i. e. emergency situations), in real-time, by analyzing the text, linked URLs, number of retweets and geographic information extracted from both post text and external URLs. Thus, an event with a higher credibility score indicates that there are more tweets, more linked URLs and more retweets mentioning this event. Data is collected through Twitter API, to get the information of the tweets and Google Maps Geocoding API to obtain geolocalization information.

In a recent context of diseases, the work presented in Yang *et al.* (2020) identifies low-credibility sources. Authors collect 570 low-credibility sources, labeled as follows: low-credibility; "Black" or "Red" or "Satire"; "fakenews" or "hyperpartisan"; or "extremeleft" or "extremeright." They use the API from the Observatory on SocialMedia to collect tweets in the Covid-19 context and the API to extract the URL, combined with regular expressions to extract any URL-like strings from the tweet text. For retweets, they include the URLs in the original tweets using the same method. They also detect bots to reduce the credibility of the tweet using BotometerLite (Yang *et al.*, 2019b).

The previous works are proposed for the English language; however, the Arabic language has been also studied. An automatically measuring the credibility of Arabic News content published in Twitter is presented in Al-Khalifa and Al-Eidan (2011). Authors select a set of features for evaluating Twitter credibility based on tweet content and author. The extraction method is the Twitter API. The features used are similarity with verified content, inappropriate words, linking to authoritative/credible, news sources, account verification, TwitterGrader.com [3] degree. The system architecture of this work consists of four main components: text pre-processing, features extraction and computation, credibility calculation and credibility assignment and ranking.

Table 2 summarizes and compares the referenced studies, in terms of the aspects taken into account to execute the credibility analysis and the extraction method used. The majority of works take into account the use of features relevant in the four elements considered as credibility measures (text, user, social and topic level), showing the variety of data needed. Most of the referenced works are developed for specific topics (Castillo *et al.*, 2011; Gupta *et al.*, 2014a; Yang *et al.*, 2020) or with a limited topic consideration such as URLs or Hashtags. However, all of them are Twitter API dependent, which is now restricted, then most of them are no longer available.

None of these studies perform a comparative evaluation of both extraction techniques; however, they are clear evidence of the need of having different options to extract different data to do credibility analysis on Twitter.

Work	Text	Credibility measures		Topic level	Web scraping	API
		User	Social			
Alrubaian *et al.* (2016b)	Yes	Yes	Yes	Sentiment	–	Yes
Al-Khalifa and Al-Eidan (2011)	Yes	Account verification	*followers*, *following*, TwitterGrade	*Hashtag*, *URLs*	–	Yes
Canini *et al.* (2010)	Yes	–	*followers*	Yes	–	Yes
Castillo *et al.* (2011)	Yes	Account verification Account creation	*followers*, *following*	Yes	–	By Twitter monitor
Lorek *et al.* (2015)	Yes	Account verification	*followers*, *following*	Location	–	Yes
Namihira *et al.* (2013)	Yes	User expertise	–	Sentiment	–	–
Yang *et al.* (2020)	Yes	–	–	URLs	–	Yes
Shao *et al.* (2016)	–	–	Yes	URLs	Yes	Yes
Gupta *et al.* (2014a)	Yes	Time of tweet	*followers*	–	–	Yes
Alrubaian *et al.* (2016a)	Yes	Time of tweet	*followers*, *following*	Hashtags	–	Yes
Liu *et al.* (2015)	Yes	Account verification	retweets	URLs, Location	Yes	Yes
Iftene *et al.* (2020)	Yes	Yes	retweets favorites	–	–	Yes
Yang *et al.* (2019a)	–	–	retweets mentions	Location URLs	–	Yes
Tan (2017)	–	–	–	Headlines URLs	Yes	Yes
Dongo *et al.* (2020)	Yes	Account verification Account creation (year)	*following* *followers*	–	Yes	Yes

Table 2.
Related work comparison

3.2 Studies about comparison

In other context different to credibility analysis, few research studies have focused on comparing both extraction techniques. With the aim of obtaining an unlimited volume of tweets, authors in Hernandez-Suarez *et al.* (2018) bypass date ranges limitations of the Twitter API, by using Web scraping, with faster results. The comparison shows that the total of retrieved tweets with the Twitter API is always less than with the Twitter Scrapy [4]. Authors establish that even though most works use the Twitter streaming API for collecting data, a limitation occurs when queries exceed rating intervals and time ranges. Authors manipulate Twitter's search query URL with the keywords they are interested in, adding also the range of dates when the tweets were made, which is a major limitation when using Twitter API. After the results page appears, the HTML payload is redirected to a Web scraping engine, where unprocessed data containing tweets is complemented to strip hypertext tags and objects, known as *tag – selectors*. Results of this study show that using this proposed methodology, more tweets are obtained in less significant seconds.

The study presented in Freelon (2018), talks about the post-API age for the next years. The author warns that even though the APIs are easy to use and TOS-compliant, they could be restricted or eliminated without warning. Web scraping has the advantage to be much more flexible but also requires more work and could evade some TOS.

The majority of works take into account the use of features relevant in the four elements considered as credibility measures (Alrubaian *et al.*, 2016a, 2016b; Al-Khalifa and Al-Eidan, 2011; Castillo *et al.*, 2011; Liu *et al.*, 2015; Lorek *et al.*, 2015), showing the variety of data needed. Most of the referenced works are developed for specific topics (Castillo *et al.*, 2011; Gupta *et al.*, 2014a; Yang *et al.*, 2020) or with a limited topic consideration such as URLs or Hashtags. However, all of them are Twitter API dependent, which is now restricted, then most of them are no longer available. Some works combine both data extraction method (Dongo *et al.*, 2019; Cardinale *et al.*, 2021; Shao *et al.*, 2016; Liu *et al.*, 2015; Tan, 2017), by using Twitter API the most of time, but when the API limitations impede the extraction of some data, then Web scraping technique is used. The limitations concern to the limited timeline, URLs access, etc. None of these studies perform a comparative evaluation of both extraction techniques, however, they are clear evidence of the need of having different options to extract different data to do credibility analysis on Twitter.

Few research has focused on comparing both extraction techniques but in other context different to credibility analysis. With the aim of obtaining an unlimited volume of tweets, authors in Hernandez-Suarez *et al.* (2018) bypass date ranges limitations of the Twitter API, by using Web scraping, with faster results. The comparison shows that the total of retrieved tweets with the Twitter API is always less than with the Twitter Scrapy [5]. Authors establish that even though most works use the Twitter streaming API for collecting data, a limitation occurs when queries exceed rating intervals and time ranges. Authors manipulate Twitter's search query URL with the keywords they are interested in, adding also the range of dates when the tweets were made, which is a major limitation when using Twitter API. After the results page appears, the HTML payload is redirected to a Web scraping engine, where unprocessed data containing tweets is complemented to strip hypertext tags and objects, known as *tag – selectors*. Results of this study show that using this proposed methodology, more tweets are obtained in less significant seconds.

The study presented in Freelon (2018), talks about the post-API age for the next years. The author warns that even though the APIs are easy to use and TOS-compliant, they could be restricted or eliminated without warning. Web scraping has the advantage to be much more flexible but also requires more work and could evade some TOS.

Beyond credibility analysis and comparison among data extraction methods, there are works that emphasize in Web scraping as an effective alternative to gather data, such as the studies presented in Glez-Peña *et al.* (2013), to extract biomedical data in real-time; in Kusumasari and Prabowo (2020) to show the pattern of use of Twitter to send warnings and identify crucial needs and responses considering Twitter as a communication channel; Kaburuan *et al.* (2019), to analyze tweets and see the commerce and tax-income potency in purposely Indonesia; and in Dewi *et al.* (2019), to propose a method able to search information, combine and present it in a better way according to user preferences.

The following section describes the credibility model proposed in Dongo *et al.* (2019).

4. Credibility analysis model for Twitter

To calculate the credibility of a tweet with Web scraping and Twitter API, we use the credibility model proposed in previous work (Dongo *et al.*, 2019). This model considers text, user and social credibilities, regardless of the topic treated. The definitions that conform to the model required in this implementation are resumed as shown below.

4.1 Text credibility

The text credibility uses syntactic analysis techniques through three *filters*:

- **SPAM Filter** (*isSPAM*): SPAM messages are usually unwanted advertising messages, which usually use hyperbolic language, excess capitalization and accentuation. If the analyzed text has these characteristics, its credibility level may decrease.
- **Bad Words Filter** (*bad_words*): It is used to detect those publications that have a high content of bad words, in which case their credibility should be reduced.
- **Misspelling Filter** (*misspelling*): It is used to detect syntax errors in writing text. Credibility will decrease as the number of misspellings found.

Based on these three filters, the text credibility is defined as shown by Def. 4.1. The user decides the importance of each filter.

Definition 4.1. Text Credibility (*TextCred*)**.** *Given a text of a tweet, t.text, the Text Credibility is a function, denoted as TextCred(t.text), that returns a measure ∈ [0,100], defined as:*

$$\begin{aligned} TextCred(t.text) = {} & w_{SPAM} \times isSpam(t.text) + \\ & w_{BadWords} \times bad\_words(t.text) + \\ & w_{MisspelledWords} \times misspelling(t.text) \end{aligned}$$

where w_{SPAM}, $w_{BadWords}$ *and* $w_{MisspelledWords}$ *represent the weights that the user gives to each filter, respectively, such as:*

$$w_{SPAM} + w_{BadWords} + w_{MisspelledWords} = 1$$

4.2 User credibility

To calculate this credibility measure on Twitter, we consider whether the account is verified or not (*Verif_Weight*) and the activity time of the account as its creation (*Creation_Weight*).

While the year the account was joined (*Year_Joined*) is closer to the *Twitter* creation date (2006), the account credibility is greater. The maximum points obtained by this criterion is

50, as the other 50 is for the verified account weight (*Verif_Weight*). Thus, the user credibility is calculated as shown in Def. 4.2.

Definition 4.2. User Credibility (*UserCred*)**.** *Given an account of a tweet, t.user, the User Credibility is a function, denoted as UserCred(t.user), that returns a measure ∈ [0,100], defined as:*

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

4.3 Social credibility

Social credibility measures the impact of a tweet on the author's social network based on his/her popularity. This credibility is calculated based on the influence of the account considering the number of *followers* (*FollowersImpact*) and the proportion of *followers* and *following* calculated by the ratio between them (*FFProportion*), each one with a maximum weight of 50. Def. 4.3 shows the calculation of social credibility.

Definition 4.3. Social Credibility (*SocialCred*)**.** *Given a set of social metadata of a tweet, t.social, the Social Credibility is a function, denoted as SocialCred(t.social), that returns a measure ∈ [0,100], defined as:*

$$SocialCred(t.social) = FollowersImpact(t.social_{user}) + FFProportion(t.social_{user})$$

4.4 Total credibility level

With the three credibility measures, the credibility level of a tweet is calculated, as shown in Def. 4.4.

Definition 4.4. Tweet Credibility Level (*TCred*)**.** *Given a tweet,* t*, the Tweet Credibility Level is a function, denoted as TCred(t), that returns a measure ∈ [0,100], of its level of credibility, defined as:*

$$TCred(t) = weight_{text} \times TextCred(t.text) + weight_{user} \times UserCred(t.user) \; \backslash\backslash \\ + weight_{social} \times SocialCred(t.social)$$

where:

- $weight_{text}$, $weigh_{user}$ *and* $weight_{social}$ *represent the weights that the user gives to text credibility, user credibility and social credibility, respectively, such as* $weight_{text} + weigh_{user} + weight_{social} = 1$*;*
- *TextCred*(*t.text*), *UserCred*(*t.user*) *and SocialCred*(*t.social*) *represent the credibility measure related to the text, the user and the social impact of* t*, respectively.*

The parameters considered in the filters and the weight that each filter represents in the final credibility calculation are values provided by users.

5. A framework for Twitter credibility analysis

In previous work, we propose a credibility model for social networks (Dongo *et al.*, 2019) (see a briefly description in Section 4). In Dongo *et al.* (2019), we also present an implementation of such as model to perform real-time credibility analysis on Twitter, based on Web scraping and implemented as a Google Chrome extension. Afterward, to make a

quantitative comparative evaluation of both data extraction methods, we extended the implementation, by updating the scraper, as the Twitter website changed its HTML tags and structure and by incorporating the use of the Twitter API, as an alternate data extraction technique (Dongo *et al.*, 2020). Thus, the current version of our framework can be configured to use Web scraping or Twitter API. In this work, we use this version of the framework to perform a qualitative and quantitative evaluation.

The framework is divided into two modules: the front-end, where the data extraction is performed in the Google Chrome extension and the back-end, which calculates the credibility. This perspective allows improving the credibility process in the future, without modifying the extension for users who already have it installed.

Figure 1 shows the dataflow of the implementation. Five features are extracted by using either Web scraping or Twitter API. Then, *Text, User* and *Social* credibility measures are calculated. The current implementation of our framework uses the following libraries, taken from a JavaScript development platform (NPM [6]): *Simple Spam Filter*, *Bad Words* and *Nspell*, for SPAM, Bad Words and Miss Spelling filters, respectively.

For the implementation of the API extraction method, we requested permissions to use the Twitter API. To obtain the API Key, a Twitter account using an institutional mail was opened, then it was registered as a developer account in the developer.twitter.com site. A form was filled to request the API Key, explaining the motives of the investigation. Once Twitter granted permissions and its API Key was obtained, the data of interest was successfully extracted:

- using user_id to obtain data from the user;
- using tweet_id to obtain data from the tweet, with an additional parameter called tweet_mode = "extended" to extract more data (i.e. an indication to not truncate tweets larger than 120 characters).

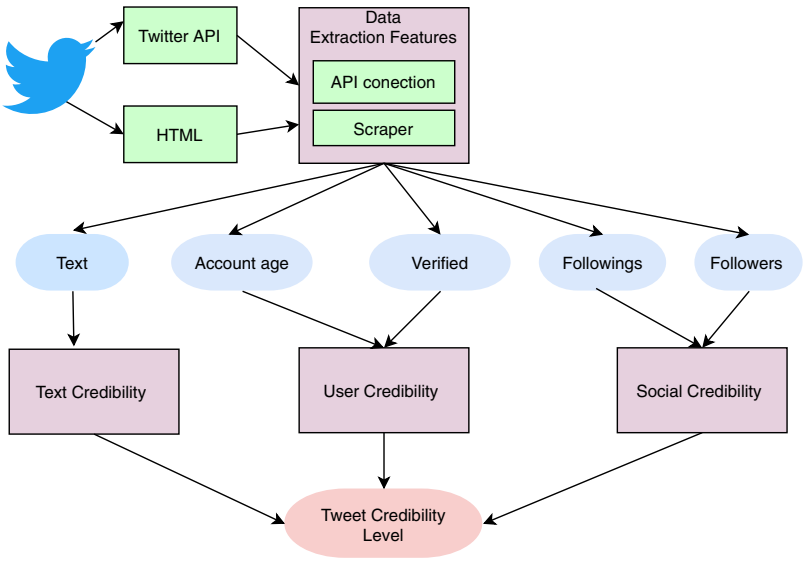

Figure 1.
Implementation of the credibility model

To obtain the user_id and tweet_id as parameters for the Twitter API, the values have to be scraped from the user profile and main homepage timelines, respectively. Table 3 shows the five attributes used to instantiate the credibility model. The simplicity of Twitter API syntax with respect to the Web scraping is clearly shown in this table; while for Twitter API, a split of the created_at value is only needed, for Web scraping, a regex and split operations are required.

The current version of our framework provides dictionaries for automatic language recognition, dictionary-en-us, dictionary-fr and dictionary-es libraries, also from NPM. Using Twitter API, we extracted the same data as the Web scraping extraction method. However, Twitter manages dates, numbers, emoticons, etc., in a different way than how it is shown on the browser, which can affect the text credibility; thus, a normalization process on the text is required.

The normalization process consists of removing and modifying some special characters using JavaScript regular expressions. First, URLs are removed from the text (tweet) and mentions, by detecting them using the regular expressions shown in Table 4. Then, hashtags and punctuation marks are located in the text and are deleted from it. Finally, to remove the emoticons, we use the library emoji-strip also provided by NPM.

6. Comparison between Web scraping and Twitter API

To compare Web scraping and Twitter API methods, we perform a battery of experiments using the credibility model proposed in our previous work (Dongo *et al.*, 2019). A qualitative and quantitative comparison are performed to analyze both methods.

Table 3. Data extraction attributes

Attributes	API	Scraper
Text	client.get('statuses/show', { id: tweetId, tweet_mode: 'extended' }).data.full_text	Array.from(document.querySelectorAll ('div [data-testid="tweet"]'))[i].children[1].children [1].children0.innerText
Verified	client.get('users/show', { user_id: userId }). data.verified	document.querySelector ('svg[aria-label="Verified account"]')
Account age	client.get('users/show', { user_id: userId }). data.created_at.split(' ').pop()	let x = document.querySelector ('div[data-testid= "UserProfileHeader_Items"]'). children[i] if x.textContent.match(/^(Joined)/) { x.textContent.split(' ')[2]}
Followings	client.get('users/show', { user_id: userId }). data.friends_count	const followingPath = window.location. pathname + '/following' document. querySelector('a[href="${followingPath}"]'). getAttribute('title')
Followers	client.get('users/show', { user_id: userId }). data.followers_count	const followersPath = window.location. pathname + '/followers' document. querySelector('a[href="${followersPath}"]'). getAttribute('title')

Table 4. Regular expressions to normalize a text

Type	Regular expression
URL	/(https?://[\^s]+)/g
Mention	/\B@[a-z0-9_-]+\s/gi
Hashtag	/#$/
Punctuation	/(|`|!|@|#|$|%||^&|*|(|)|{|}|[|]|;|:|"|'|¡|,|.|¿|?|/|||-|_|+|=)/g

6.1 Qualitative comparison

Several measures have been proposed in the literature to evaluate the tweet' credibility but due to the complexity of measures and the availability of attributes, some measures are not possible to implement for an automatic calculation. This qualitative comparison describes the attributes that can be retrieved and therefore the measures that can be performed using Web scraping and Twitter API.

The qualitative tests are described as follows:

- Test 1 is to compare the attributes that can be retrieved by using Web scraping and the Twitter API.
- Test 2 is intended to describe which credibility measures available in the literature can be performed by using the attributes extracted for both methods.

6.1.1 Test 1: Available attributes. Several Twitter' attributes are used to calculate credibility measures; however, not all attributes related to a tweet are shown on the Twitter page. For instance, the attribute *truncated*, which indicates whether the value of the text parameter was truncated, for example, as a result of a retweet exceeding the original Tweet text length limit of 140 characters, can be retrieved using Twitter API, while using Web scraping, it is not possible. Table 5 shows a list of some Twitter' attributes, their description, data type and availability for both Web scraping and Twitter API extraction methods. In general, only 54.29% of the attributes listed in the table can be retrieved using Web scraping.

6.1.2 Test 2: Credibility measures. Using Twitter' attributes, several credibility measures are performed; therefore, if some attributes cannot be obtained by Web scraping or Twitter API, the measures cannot be calculated automatically. Authors in Riquelme and González-Cantergiani (2016) perform a study on the influence of Twitter and describe several credibility measures. Based on this study, we present in Table 6 some credibility measures and their availability according to the extraction methods. In total, 60% of the measures can be performed by Twitter API, while only 40% for Web scraping.

6.2 Quantitative comparison

In this section, a quantitative comparison is performed to discover some differences in the credibility values between Web scraping and Twitter API by an empirical evaluation. Additionally, the requesting time of the Twitter API is also evaluated. The tests are described as follows:

- *Test 1* is intended to analyze the use of special characters, which can affect credibility (in Text Credibility).
- As the User Credibility only takes into account the verified account and joined date and the obtained values are identical for both extraction methods, their credibilities are the same and do not merit an evaluation.
- *Test 2* is to study Social Credibility, where two different Twitter accounts (famous and common) are considered for a better analysis.
- *Test 3* is to apply the whole credibility model to four tweets of each Twitter account, evaluated in *Test 2*.
- Test 4 is to evaluate the requesting time of using the Twitter API from different locations around the world.
- *Test 5*, the past experiment, is to compare the execution time of the two extraction methods.

Table 5. Some Twitter' attributes

Attribute	Sub-attribute	Description	Data type	Example	Twitter API	Web Scraping
created_at	–	UTC time when this Tweet was created	String	"Wed Oct 10 20:19:24 + 0000 2018"	Yes	Yes
id	–	The integer representation of the unique identifier for this Tweet	Int64	1050118621198921728	Yes	Yes
id_str	–	The string representation of the unique identifier for a Tweet	String	"1050118621198921728"	Yes	No
text	–	The actual UTF-8 text of the status update	String	"To make room for more ex[. . .]"	Yes	Yes
source	–	Utility used to post the Tweet, as an HTML-formatted string	String	"Twitter Web Client"	Yes	Yes
truncated	–	Indicates whether the value of the text parameter was truncated, for example, as a result of a retweet exceeding the original Tweet text length limit of 140 characters	Boolean	true	Yes	No
in_reply_to_status_id	–	Nullable. If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID	Int64	1051222721923756032	Yes	No
in_reply_to_status_id_str	–	Nullable. If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's ID	String	"1051222721923756032"	Yes	No
in_reply_to_user_id	–	Nullable. If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's author ID	Int64	6253282	Yes	No
in_reply_to_user_id_str	–	Nullable. If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's author ID	String	"6253282"	Yes	No
in_reply_to_screen_name	–	Nullable. If the represented Tweet is a reply, this field will contain the screen name of the original Tweet's author	String	"twitterapi"	Yes	Yes
user	id	The integer representation of the unique identifier for this user	Int64	6253282	Yes	Yes
	id_str	The string representation of the unique identifier for this user	String	"6253282"	Yes	No
	name	The name of the user, as they have defined it. Not necessarily a person's name	String	"Twitter API"	Yes	Yes
	screen_name	The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change	String	"twitterapi"	Yes	Yes
	location	Nullable. The user-defined location for this account's profile. Not necessarily a location, nor machine-parseable	String	"San Francisco, CA"	Yes	Yes
	url	Nullable. A URL provided by the user in association with their profile	String	"https://developer.twitter.com"	Yes	Yes

(*continued*)

Table 5.

Attribute	Sub-attribute	Description	Data type	Example	Twitter API	Web Scraping
	description	Nullable. The user-defined UTF-8 string describing their account.	String	"The Real Twitter API."	Yes	No
	verified	When true, indicates that the user has a verified account	Boolean	False	Yes	Yes
	followers_count	The number of followers this account currently has	Int	21	Yes	Yes
	friends_count	The number of users this account is following (AKA their "followings")	Int	32	Yes	Yes
	listed_count	The number of public lists that this user is a member of	Int	9,274	Yes	No
	favourites_count	The number of Tweets this user has liked in the account's lifetime	Int	13	Yes	Yes
	statuses_count	The number of Tweets (including retweets) issued by the user	Int	42	Yes	Yes
	created_at	The UTC datetime that the user account was created on Twitter	String	"Mon Nov 29 21:18:15 + 0000 2010"	Yes	Yes
	utc_offset	The value will be set to null. Still available via GET account/settings	null	–	Yes	No
	time_zone	The value will be set to null. Still available via GET account/settings as tzinfo_name	null	–	Yes	No
	geo_enabled	The value will be set to null. Still available via GET account/settings	null	–	Yes	No
	lang	The value will be set to null. Still available via GET account/settings as language	null	–	Yes	Yes
	profile_image_url_https	A HTTPS-based URL pointing to the user's profile image	String	"https://abs.twimg.com/...png"	Yes	Yes
	profile_banner_url	The HTTPS-based URL pointing to the standard Web representation of the user's uploaded profile banner	String	"https://si0.twimg.com/...824"	Yes	Yes
	default_profile	When true, indicates that the user has not altered the theme or background of their user profile	Boolean	False	Yes	No
	default_profile_image	When true, indicates that the user has not uploaded their own profile image and a default image is used instead	Boolean	False	Yes	No
	following	The value will be set to null. Still available via GET friendships/lookup	null	–	Yes	Yes
	follow_request_sent	The value will be set to null. Still available via GET friendships/lookup	null	–	Yes	Yes

Metric	Description	Twitter API	Web scraping
Tweet count score (Noro *et al.*, 2012)	Counts the number of original tweets plus the number of retweets	Yes	No
Signal strength (Pal and Counts, 2011)	How strong is the author's topical signal; it measures the originality of the author's tweets	Yes	No
Effective readers (Lee *et al.*, 2010)	Sum of effective readers for all user tweets; where an effective reader of a tweet is a follower who still has not tweeted on any trending topic when the user sent the tweet	Limited request	No
ActivityScore (Yuan, 2013)	This measure counts the number of followers, following and tweets on a graph for each user during a period of time	Limited request	No
DiscussRank (Ben Jabeur *et al.*, 2012)	It determines how active a user is, in the sense of initiating conversations around a topic	Limited request	No
FollowerRank (Nagmoti *et al.*, 2010)	It is the normalized version of the traditional in-degree measure for social networks	Yes	Yes
Popularity (Aleahmad *et al.*, 2015)	It is the level of popularity based on the number of followers of the user	Yes	Yes
IP Influence (Romero *et al.*, 2011)	It measures the users' influence and he/she passivity. The passivity of a user is defined as the difficulty for the user to be influenced by another in some period of time	Limited request	No
Retweet Impact (Pal and Counts, 2011)	It estimates the impact of the user tweets, in terms of the retweeted tweets	Yes	Yes
Mention Impact (Pal and Counts, 2011)	It estimates the impact of the user tweets, in terms of the mentions received by other users	Yes	Yes

Table 6.
Credibility measures

Tests 1, 2 and *3* were applied to three different Twitter account languages: Spanish, English and French. These tests and *Test 5* were undertaken in a shared VPS on Digital Ocean with 1 GB Memory, 20 GB Disk and Ubuntu 14.04.4 ×64. The VPS is located in New York City. The values of the features correspond to the date of July 28, 2020.

Test 4 was undertaken in various shared VPS on Digital Ocean with 1 GB Memory, 25 GB Disk, and Ubuntu 20.04 ×64, located in San Francisco, Toronto, New York, London, Amsterdam, Singapore and Bangalore. The results correspond to the date of March 18, 2021.

6.2.1 Test 1: Text credibility. To study the impact of text retrieval using Web scraping and Twitter API, we consider scenarios where different types of tweets such as short, long, use of SPAM, bad words and miss spelling and emoticons, are presented.

Table 7 shows the results obtained for the Spanish tweets. As a normalized process is applied to the text obtained by Web scraping and Twitter API, consisting of removing links, emoticons and other steps (see more details in Section 5), the Text Credibility values are the same for both extraction methods. A similar scenario can be observed for English and

Type	Twitter	Tweet ID	Info	Extraction method	SPAM (%)	Bad words (%)	Miss spelling (%)	Text credibility (%)
Short	@yuniquintero	X...792	12 words	Web scraping	100	100	16.6666	72.5
				Twitter API	100	100	16.6666	72.5
Long	@nanutria	X...192	43 words	Web scraping	100	100	62.9629	95.7692
				Twitter API	100	100	62.9629	95.7692
SPAM	@farmarato	X...360	2 words	Web scraping	0	100	0	33
				Twitter API	0	100	0	33
Bad words	@yuniquintero	X...240	16 words 1 emoticon	Web scraping	100	94.1176	17.6470	70.8823
				Twitter API	100	94.1176	100	98.0588
Misspelling	@eldtwiter	X...857	5 words	Web scraping	100	100	60	86.8
				Twitter API	100	100	60	86.8
Emoticons	@fabi_ad	X...079	22 words 2 emoticons	Web scraping	100	100	94.7368	98.2631
				Twitter API	100	100	94.7368	98.2631

Note: Text credibility = 0.34 × SPAM + 0.33 × Bad words + 0.33 × Misspelling

Table 7. *Test 1*: Text credibility – Spanish

French tweets, Tables 8 and 9, respectively. Moreover, a short or long text does not have an impact on credibility. For example, the credibility value for long-type text is bigger than the one of the short-type text (95.7692% and 72.5%, respectively), in the case of the Spanish language, while for English, the credibility value of the long-type text is less than the one of short-type text (95.7692% and 100%, respectively). *Bad words* and *Misspelling* filters provided low credibility values for the types related to them, which proves their functionality.

6.2.2 Test 2: Social credibility. For this test, we select two Twitter accounts for each language, taking into account the number of *followers* (i.e. common and famous accounts). For the Spanish language, we selected @YuniQuintero and @presidenciaperu; for English, @chen_bichan and @elonmusk; while for French, @cocopericau and @antogriezmann. A value of 2 million is used for the *Max Followers* parameter, as we proposed in Dongo *et al.* (2019). Table 10 shows the results obtained for this test. Similarly to *Test 1*, the results obtained for Web scraping and Twitter API are the same. However, there is a small difference of *followers* between Web scraping and Twitter API, for famous accounts, as this number is constantly increasing and the tests were performed consecutively (one after the

Type	Twitter	Tweet ID	Info	Extraction method	SPAM (%)	Bad words (%)	Miss spelling (%)	Text credibility (%)
Short	@elonmusk	X...427	6 words	Web scraping	100	100	100	100
				Twitter API	100	100	100	100
Long	@elonmusk	X...142	37 words	Web scraping	100	100	97.2222	99.0833
				Twitter API	100	100	97.2222	99.0833
SPAM	@maheshfantrends	X...601	14 words	Web scraping	0	100	86.6666	61.6
				Twitter API	0	100	86.6666	61.6
Bad Words	@chen_bichan	X...944	6 words	Web scraping	0	83.3333	100	60.5
				Twitter API	0	83.3333	100	60.5
Misspelling	@gitlost	X...024	3 words 14 emoticons	Web scraping	0	66.6666	33.3333	32.9999
				Twitter API	0	66.6666	33.3333	32.9999
Emoticons	@elonmusk	X...987	2 words 1 emoticon	Web scraping	0	100	100	66
				Twitter API	0	100	100	66

Note: Text credibility = 0.34 × SPAM + 0.33 × Bad Words + 0.33 × Misspelling

Table 8. *Test 1*: Text credibility – English

Type	Twitter	Tweet ID	Info	Extraction method	SPAM (%)	Bad words (%)	Miss spelling (%)	Text credibility (%)
Short	@_rapvibes	X...664	8 words	Web scraping	100	100	100	100
				Twitter API	100	100	100	100
Long	@jml_932	X...011	43 words	Web scraping	0	100	28.8888	42.5333
				Twitter API	0	100	28.8888	42.5333
SPAM	@opcrotte	X...473	8 words	Web scraping	0	100	56.25	51.56
				Twitter API	0	100	56.25	51.56
Bad words	@_rapvibes	X...482	21 words	Web scraping	100	95.2390	90.4761	95.2857
				Twitter API	100	95.2390	90.4761	95.2857
Misspelling	@cocopericaud	X...129	33 words	Web scraping	100	100	91.1764	97.9705
				Twitter API	100	100	91.1764	97.9705
Emoticons	@antogriezmann	X...067	5 words 2 emoticons	Web scraping	0	100	100	66
				Twitter API	0	100	100	66

Table 9.
Test 1: Text credibility – French

Notes: Text credibility = 0.34 × SPAM + 0.33 × Bad words + 0.33 × Misspelling

Type	Twitter	Max Followers	Extraction method	Followers	Followings	Followers Impact (%)	FF Proportion (%)	Social Credibility (%)
Spanish	@yuniquintero	2 M	Web scraping	411	266	0.0204	60.7090	30.3647
			Twitter API	411	266	0.0204	60.7090	30.3647
	@presidenciaperu	2 M	Web scraping	933,099	277	46.6548	99.9702	73.3126
			Twitter API	933,100	277	46.6548	99.9702	73.3126
English	@chen_bichan	2 M	Web scraping	368	851	0.0184	30.1886	15.1035
			Twitter API	368	851	0.0184	30.1886	15.1035
	@elonmusk	2 M	Web scraping	37,172,852	97	100	100	100
			Twitter API	37,172,865	97	100	100	100
French	@cocopericaud	2 M	Web scraping	571	696	0.0284	43.0670	22.5478
			Twitter API	571	696	0.0284	43.0670	22.5478
	@antogriezman	2 M	Web scraping	7,015,909	10	100	100	100
			Twitter API	7,015,914	10	100	100	100

Table 10.
Test 2: Social credibility analysis

Note: Social credibility = 0.50 × Followers impact + 0.50 × FF proportion

other). For example, for @elonmusk, the number of *followers* for Web scraping was 37'172,852, while for Twitter API was 37'172,865. Note that the Web scraping method was performed before its respective Twitter API test.

6.2.3 Test 3: Tweet credibility. By using the six previous Twitter accounts, we select randomly four of the most recent tweets for each account. We calculate the tweet credibility and we report the text, user and social credibilities. Tables 11, 12, and 13 show the results for Spanish, English and French, respectively. As previous tests, Web scraping and Twitter API extractions methods produce the same credibility values. Moreover, we can observe that famous accounts have better credibility than the common ones due to their user and social credibilities. The social credibility for a famous account is around 100% avg., while for the other ones is 22% avg. The famous accounts are all verified which has a huge impact on the total credibility (16.67%).

Tweet ID	Extraction method	Total credibility (%)	Text credibility (%)	User credibility (%)	Social credibility (%)
@YuniQuintero					
XXXXX29977786449921	Web scraping	43.8218	61.2857	39.2857	30.3647
	Twitter API	56.9846	100	39.2857	30.3647
XXXXX81616200945665	Web scraping	41.3969	54.1538	39.2857	30.3647
	Twitter API	56.9846	100	39.2857	30.3647
XXXXX57961848909826	Web scraping	41.9723	55.8461	39.2857	30.3647
	Twitter API	56.9846	100	39.2857	30.3647
XXXXX72551944265728	Web scraping	41.6846	55	39.2857	30.3647
	Twitter API	34.2046	49.5	39.2857	30.3647
@presidenciaperu					
XXXXX44070680453122	Web scraping	78.2113	75.6842	85.7142	73.3122
	Twitter API	83.8895	92.3846	85.7142	73.3122
XXXXX13251207360514	Web scraping	78.5426	76.6585	85.7142	73.3122
	Twitter API	84.6087	94.5	85.7142	73.3122
XXXXX92509447139328	Web scraping	78.0637	75.25	85.7142	73.3122
	Twitter API	85.0762	95.875	85.7142	73.3122
XXXXX91373758885891	Web scraping	78.2911	75.9189	85.7142	73.3122
	Twitter API	83.8213	92.1842	85.7142	73.3122

Note: Total credibility = 0.34 × Text credibility + 0.33 × User credibility + 0.33 × Social credibility

Table 11.
Test 3: Max followers: 2 M – Spanish

Tweet ID	Extraction method	Total credibility (%)	Text credibility (%)	User credibility (%)	Social credibility (%)
@chen_bichan					
XXXXX97846390792192	Web scraping	39.9441	95.875	7.1428	15.1197
	Twitter API	39.9441	95.875	7.1428	15.1197
XXXXX62917602504705	Web scraping	41.3466	100	7.1428	15.1197
	Twitter API	41.3466	100	7.1428	15.1197
XXXXX61693427703810	Web scraping	40.3266	97	7.1428	15.1197
	Twitter API	40.3266	97	7.1428	15.1197
XXXXX60908505718784	Web scraping	40.1859	96.5862	7.1428	15.1197
	Twitter API	40.1859	96.5862	7.1428	15.1197
@elonmusk					
XXXXX75982297051142	Web scraping	95.6628	97.6428	89.2857	100
	Twitter API	95.6628	97.6428	89.2857	100
XXXXX69404874088448	Web scraping	84.9042	66	89.2857	100
	Twitter API	84.9042	66	89.2857	100
XXXXX91044466724866	Web scraping	95.5292	97.25	89.2857	100
	Twitter API	95.5292	97.25	89.2857	100
XXXXX77935580160000	Web scraping	96.4642	100	89.2857	100
	Twitter API	96.4642	100	89.2857	100

Note: Total credibility = 0.34 × Text credibility + 0.33 × User credibility + 0.33 × Social credibility

Table 12.
Test 3: Max followers: 2 M – English

6.2.4 Test 4: Twitter API performance. Twitter API provides several methods to obtain Twitter' attributes for different purposes such as develop of applications and data analysis. This API can be accessed from different locations around the world. In this test, we evaluated the performance of requesting time by calling the Twitter API from San

Tweet ID	Extraction method	Total credibility (%)	Text credibility (%)	User credibility (%)	Social credibility (%)
@cocopericaud					
XXXXX72931879002112	Web scraping	45.6120	94.9230	17.8571	22.5617
	Twitter API	45.6120	94.9230	17.8571	22.5617
XXXXX11753465520129	Web scraping	46.648174	97.9705	17.8571	22.5617
	Twitter API	46.648174	97.9705	17.8571	22.5617
XXXXX08508903104512	Web scraping	30.2209	49.6551	17.8571	22.5617
	Twitter API	30.2209	49.6551	17.8571	22.5617
XXXXX67680429191169	Web scraping	45.4682	94.5	17.8571	22.5617
	Twitter API	45.4682	94.5	17.8571	22.5617
@antogriezman					
XXXXX55081208217602	Web scraping	80.9370	64.7307	78.5714	100
	Twitter Api	80.9370	64.7307	78.5714	100
XXXXX91677961048067	Web scraping	81.3685	66	78.5714	100
	Twitter Api	81.3685	66	78.5714	100
XXXXX75192480796684	Web scraping	72.9535	41.25	78.5714	100
	Twitter Api	72.9535	41.25	78.5714	100
XXXXX04159384465408	Web scraping	81.3685	66	78.5714	100
	Twitter Api	81.3685	66	78.5714	100

Note: Total credibility = 0.34 × Text credibility + 0.33 × User credibility + 0.33 × Social credibility

Table 13. *Test 3*: Max followers: 2 M – French

Francisco, Toronto, New York, London, Amsterdam, Singapore and Bangalore. The average of 10 executions are reported in Table 14. As Twitter is located at San Francisco, calls from that same city have the best performance (44.2372 ms), while Singapore and Bangalore have 195.3096 ms and 319.1896 ms, respectively. The results show a direct proportion in terms of performance and distance, i.e. the more distance, the more requesting time. Some places as Toronto and Singapore do not follow the direct proportion rule due to the network connection between San Francisco and these cities.

6.2.5 Test 5: Performance. Finally, we measured the processing time of the two extraction methods, from the initial request until all five features for the credibility model are obtained. The average of 10 repetitions for each tweet is reported in Table 15. Results show that Web scraping is 40 times faster than Twitter API as to obtain the user_id and tweet_id, Web scraping is also used; thus, a Twitter API call consists of local processing (Web scraping) and an API request.

The following section discusses about the differences between the two techniques, Web scraping and Twitter API. Furthermore, the results obtained of the performed tests are also analyzed.

Location	Distance (km)	Time (ms)
San Francisco	0	44.2372
Toronto	3,634.01	130.1669
New York	4,121.40	105.1686
London	8,608.48	173.5531
Amsterdam	8,763.78	173.2388
Singapore	13,595.30	195.3096
Bangalore	13,968.32	319.1896

Table 14. Execution time par location

6.3 Discussion

Data extraction methods have been widely used in social networks, in special on Twitter. When there are API limitations, some works come up with alternates and bypass the APIs, using Web scraping to gather the data needed.

Web scraping is more flexible than API extraction because it can be used on most Web pages, not just those that offer APIs (Freelon, 2018). Web scraping would not be necessary if each website provides an API to share their data in a structured format. However, some websites have APIs, but they are restricted by what data is available and how frequently it can be accessed (Dongo *et al.*, 2019). For Web scraping, the speed can affect the data extraction when the page is not displayed, but its use is free for some cases where policies allow. Moreover, the constant change of the Web pages affects this method. With APIs, the access to extract data is limited and it is not free, however, its use is independent of the information displayed on the website. Then, we can observe clearly two main differences between Web scraping and APIs, the speed of getting the data (network connection) and the access (number of requests). The variety of the products and the information that can be extracted from the API enables to have a question whether APIs can be a feasible method of extracting data posted on the Web.

By a qualitative evaluation, *Test 1* and *Test 2* demonstrated that more attributes and therefore more credibility measures can be performed using Twitter API than Web scraping. Twitter API provides many attributes related to the tweet, user, location, etc., which makes the different with Web scraping.

Regardless of the method used, we assume that the data obtained should be similar as there must be a consistency between the displayed information (Web scraping) for users and the one obtained by the API, for developers. To affirm or deny our assumption, we evaluated, using our proposed model, the tweet credibility in real-time for both methods, Web scraping and Twitter API.

The scraper developed for our previous work (Dongo *et al.*, 2019), did not work when we started this study because of changes in the Twitter HTML. Then, we developed another one following the new HTML tags, to perform the comparative evaluation. After finished the experiments, we wanted to validate some of the results, but the Twitter HTML format changed again and we could not do it. This is the main disadvantage of Web scraping, data can be easily accessed but the extraction functions have to follow the current HTML tags.

On the other hand, Twitter API manages dates, numbers and emoticons in a different way than how they are shown in the browser; however, by applying a normalization process on the text, we are able to obtain identical texts and thus, the same text credibility values. This is one of the most important steps during the extraction of data, as the text is more complex than the other aspects that matter for credibility (e.g. *verified account* is just a boolean value, *account age* is just an integer, see Table 3) due to the presence of special characters.

Table 15. Extraction time comparison

Twitter	Tweet ID	Extraction method	Time (ms)
@equipedefrance	XXXXX66822152019968	Web scraping	4.18
		Twitter API	222.81
@elonmusk	XXXXX71909906702336	Web scraping	5.74
		Twitter API	203.25
@presidenciaperu	XXXXX66760141852673	Web scraping	5.76
		Twitter API	199.56

In the quantitative evaluation, *Test 1* shows the effectiveness of the text normalization process. The text credibility values are identical for both extraction methods. For *Test 2*, a minor different among the number of *followers*, for a famous account, can be observed but irrelevant for the final score. This behavior is caused by the rapid and constant growth of the *followers* through time. For example, @elonmusk had 37'172,852 for Web scraping and 37'172,865 for Twitter API. *Test 3* confirms the previous tests and the total credibility confirmed by the text, user and social credibilities, are identical for both extraction methods.

Test 4 showed a faster API response when the client is located next to the Twitter location company. Some network connections can improve access time as in the case of Singapore. Finally, *Test 5* demonstrated that Web scraping is faster than Twitter API as the latter is composed by a Web scraping process to obtain the user_id and tweet_id and the API call.

7. Conclusions and future work

In this paper [7], we present a comparison between two data extraction techniques on the Web, Web scraping and API, by using an existing real-time credibility model applied to Twitter for a quantitative evaluation. To do so, we implemented a framework as an extension of Google Chrome consisting of a front-end and a back-end, which performs the credibility analysis in real-time and can be configured to use either Web scraping or Twitter API to gather the needed data to feed the credibility model. The credibility model, previously proposed in Dongo *et al.* (2019), computes a post's credibility based on Text credibility, User credibility and Social credibility.

In the qualitative evaluation, results show that more credibility measures can be performed using Twitter API than Web scraping due to the availability of attributes. In the case of the quantitative evaluation, results show that a robust normalization process on the text obtained by the extraction methods produces identical credibility results. Moreover, the number of *followers* obtained by the extraction methods have a minor different for famous accounts as the number of followers is constantly growing. The requesting time is less when a client is located in the Twitter location company (San Francisco) and this increases according to the distance. Additionally, Web scraping is faster than Twitter API as for this latter, the use of Web scraping is required to obtain the user_id and tweet_id before to perform the API request.

We are currently working on improvement the credibility model, by considering Topic credibility and extending Text, User and Social credibility with other data, such as semantic analysis, retweets and bot detection.

Notes

1. https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media
2. The Twitter monitor was an on-line monitoring system which detected sharp increases ("bursts") in the frequency of sets of keywords found in messages.
3. Twittergrader.com is a service to retrieve the grade of any Twitter user via its username.
4. Twitter Scrapy is an open source and collaborative framework for extracting data from websites. https://scrapy.org/
5. Twitter Scrapy is an open source and collaborative framework for extracting data from websites. https://scrapy.org/
6. https://www.npmjs.com/
7. This work is an extension of a paper accepted in iiWAS 2020 conference.

References

Aleahmad, A., Karisani, P., Rahgozar, M. and Oroumchian, F. (2015), "Olfinder: finding opinion leaders in online social networks", *Journal of Information Science*, Vol. 42.

Al-Khalifa, H. and Al-Eidan, R. (2011), "An experimental system for measuring the credibility of news content in twitter", *International Journal of Web Information Systems*, Vol. 7 No. 2, pp. 130-151.

Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. and Alamri, A. (2016a), "Credfinder: a real-time tweets credibility assessing system", *International Conference on Advances in Social Networks Analysis and Mining*, pp. 1406-1409.

Alrubaian, M., Al-Qurishi, M., Hassan, M. and Alamri, A. (2016b), "A credibility analysis system for assessing information on twitter", *IEEE Transactions on Dependable and Secure Computing*, Vol. 15 No. 4, pp. 661-674.

Alrubaian, M., Al-Qurishi, M., Alamri, A., Al-Rakhami, M., Hassan, M. and Fortino, G. (2019), "Credibility in online social networks: a survey", *IEEE Access*, Vol. 7, pp. 2828-2855.

Ben Jabeur, L., Tamine, L. and Boughanem, M. (2012), "Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks", in Calderón-Benavides, L., González-Caro, C., Chávez, E. and Ziviani, N. (Eds), *String Processing and Information Retrieval*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 111-117.

Boillot, M. (2012), "Application programming interface (API) for sensory events", US Patent 8,312,479.

Bovet, A. and Makse, H. (2019), "Influence of fake news in twitter during the 2016 us presidential election", *Nature Communications*, Vol. 10 No. 1.

Canini, K., Suh, B. and Pirolli, P. (2010), "Finding relevant sources in twitter based on content and social structure", *NIPS Workshop*.

Cardinale, Y., Dongo, I., Robayo, G., Cabeza, D., Aguilera, A. and Medina, S. (2021), "T-creo: a twitter credibility analysis framework", *IEEE Access*, Vol. 9, pp. 32498-32516.

Castillo, C., Mendoza, M. and Poblete, B. (2011), "Information credibility on Twitter", *International conference on WWW*, pp. 675-684.

Dewi, L., Meiliana. and Chandra, A. (2019), "Social media web scraping using social media developers API and regex", *Procedia Computer Science*, Vol. 157, pp. 444-449.

Dig (2020), "Digital 2020: 3.8 billion people use social media - we are social", [Online; accessed 18. Jul. 2020].

Dongo, I., Cardinale, Y. and Aguilera, A. (2019), "Credibility analysis for available information sources on the web: a review and a contribution", *4th International Conference on System Reliability and Safety (ICSRS)*, pp. 116-125.

Dongo, I., Cadinale, Y., Aguilera, A., Martínez, F., Quintero, Y. and Barrios, S. (2020), "Web scraping versus twitter API: a comparison for a credibility analysis", *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications and Services, iiWAS '20*, Association for Computing Machinery, New York, NY, pp. 263-273.

Edgerly, S. and Vraga, E. (2019), "The blue check of credibility: does account verification matter when evaluating news on twitter?", *Cyberpsychology, Behavior, and Social Networking*, Vol. 22 No. 4, pp. 283-287.

Freelon, D. (2018), "Computational research in the Post-API age", *Political Communication*, Vol. 35 No. 4, pp. 665-668.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. and Fdez-Riverola, F. (2013), "Web scraping technologies in an API world", *Briefings in Bioinformatics*, Vol. 15 No. 5, pp. 788-797.

Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P. (2014a), *TweetCred: Real-Time Credibility Assessment of Content on Twitter*, Springer International Publishing, Cham, pp. 228-243.

Gupta, A., Lamba, H., Kumaraguru, P. and Joshi, A. (2014b), "Analyzing and measuring the spread of fake content on twitter during high impact events", *Security and Privacy Symposium 2014, CSE-IIT-Kanpur*.

Gupta, S., Sachdeva, S., Dewan, P. and Kumaraguru, P. (2018), "CBI: improving credibility of user-generated content on Facebook", in Mondal, A., Gupta, H., Srivastava, J., Reddy, P.K. and Somayajulu, D. (Eds), *Big Data Analytics*, Springer International Publishing, Cham, pp. 170-187.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V. and Perez-Meana, H. (2018), "A web scraping methodology for bypassing twitter API restrictions", *Computing Research Repository (CoRR)*, arXiv.

Iftene, A., Gîfu, D., Miron, A. and Dudu, M. (2020), "A real-time system for credibility on twitter", *12th Language Resources and Evaluation Conference*, pp. 6166-6173.

Kaburuan, E.R., Lindawati, A.S.L., Putra, M.R. and Utama, D.N. (2019), "A model configuration of social media text mining for projecting the online-commerce transaction (case: Twitter tweets scraping)", *7th International Conference on Cyber and IT Service Management (CITSM)*, Vol. 7, pp. 1-4.

Kusumasari, B. and Prabowo, N. (2020), "Scraping social media data for disaster communication: how the pattern of twitter users affects disasters in Asia and the pacific", *Natural Hazards*, Vol. 103 No. 3.

Lee, C., Kwak, H., Park, H. and Moon, S. (2010), "Finding influentials based on the temporal order of information adoption in twitter", *Proceedings of the 19th international conference on World wide web*, pp. 1137-1138.

Liu, X., Nourbakhsh, A., Li, Q., Fang, R. and Shah, S. (2015), "Real-time rumor debunking on twitter", *International conference on Information and Knowledge Management*, pp. 1867-1870.

Lorek, K., Suehiro-Wiciński, J., Jankowski-Lorek, M. and Gupta, A. (2015), "Automated credibility assessment on twitter", *Computer Science*, Vol. 16 No. 2.

Mitchell, R. (2015), *Web Scraping with Python: Collecting Data from the Modern Web*, 1st ed., O'Reilly Media.

Nagmoti, R., Teredesai, A. and De Cock, M. (2010), "Ranking approaches for microblog search", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 153-157.

Namihira, Y., Segawa, N., Ikegami, Y., Kawai, K., Kawabe, T. and Tsuruta, S. (2013), "High precision credibility analysis of information on Twitter", *International Conference on Signal-Image Technology and Internet-Based Systems*, pp. 909-915.

Noro, T., Ru, F., Xiao, F. and Tokuda, T. (2012), "Twitter user rank using keyword search", *Information Modelling and Knowledge Bases XXIV. Frontiers in Artificial Intelligence and Applications*, Vol. 251.

Pal, A. and Counts, S. (2011), "Identifying topical authorities in microblogs", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*, Association for Computing Machinery, New York, NY, pp. 45-54.

Riquelme, F. and González-Cantergiani, P. (2016), "Measuring user influence on twitter", *Information Processing and Management*, Vol. 52 No. 5, pp. 949-975.

Romero, D.M., Galuba, W., Asur, S. and Huberman, B.A. (2011), "Influence and passivity in social media", in Gunopulos, D., Hofmann, T., Malerba, D. and Vazirgiannis, M. (Eds), *Machine Learning and Knowledge Discovery in Databases*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 18-33.

Salt, D. and Sellhorn, A. (2014), "Method, system and computer program product for a client application programming interface (API) in a service oriented architecture", US Patent 8,701,128.

Sánchez-Rada, J. and Iglesias, C. (2019), "Social context in sentiment analysis: formal definition, overview of current trends and framework for comparison", *Information Fusion*, Vol. 52, pp. 344 -356.

Shao, C., Ciampaglia, G., Flammini, A. and Menczer, F. (2016), "Hoaxy: a platform for tracking online misinformation", 25th international conference Companion on World Wide Web, pp. 745-750.

Slamet, C., Andrian, R., Maylawati, D., Suhendar, Darmalaksana, W. and Ramdhani, M. (2018), "Web scraping and naïve bayes classification for job search engine", *IOP Conference Series: Materials Science and Engineering*, Vol. 288, p. 012038.

Tan, S. (2017), "Spot the lie: detecting untruthful online opinion on twitter", Master Thesis, Department of Computing, Imperial College London.

Vaidya, T., Votipka, D., Mazurek, M. and Sherr, M. (2019), "Does being verified make you more credible?: account verification's effect on tweet credibility", *Conference on Human Factors in Computing Systems*, pp. 1-13.

Yang, J., Yu, M., Qin, H., Lu, M. and Yang, C. (2019a), "A twitter data credibility framework-hurricane Harvey as a use case", *ISPRS International Journal of Geo-Information*, Vol. 8 No. 3, p. 111.

Yang, K. Varol, O. Davis, C. Ferrara, E. Flammini, A. and Menczer, F. (2019b), "Arming the public with AI to counter social bots", CoRR, abs/1901.00912.

Yang, K. Torres-Lugo, C. and Menczer, F. (2020), "Prevalence of low-credibility information on twitter during the covid-19 outbreak", *arXiv preprint arXiv:2004.14484*.

Yuan, J. (2013), "Topology-based algorithm for users' influence on specific topics in micro-blog", *Journal of Information and Computational Science*, Vol. 10 No. 8, pp. 2247-2259.

Zannettou, S., Sirivianos, M., Blackburn, J. and Kourtellis, N. (2019), "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans", *Journal of Data and Information Quality*, Vol. 11 No. 3, pp. 1-37.

Corresponding author
Ana Aguilera can be contacted at: ana.aguilera@uv.cl