Expert Systems With Applications 164 (2021) 114006

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks☆

Despoina Antonakaki [b,*], Paraskevi Fragopoulou [b], Sotiris Ioannidis [a,b]

[a] Technical University of Crete, University Campus, Akrotiri, Chania 73100, Crete, EL 090034024, Greece
[b] Institute of Computer Science (ICS) of the Foundation for Research and Technology - Hellas (FORTH), N. Plastira 100 Vassilika Vouton, GR-700 13 Heraklion, Crete, EL 090101655, Greece

ARTICLE INFO

Keywords:
Social networks
Twitter
Survey
Social graph
Sentiment analysis
Spam
Bots
Fake news
Hate speech

ABSTRACT

Twitter is the third most popular worldwide Online Social Network (OSN) after Facebook and Instagram. Compared to other OSNs, it has a simple data model and a straightforward data access API. This makes it ideal for social network studies attempting to analyze the patterns of online behavior, the structure of the social graph, the sentiment towards various entities and the nature of malicious attacks in a vivid network with hundreds of millions of users. Indeed, Twitter has been established as a major research platform, utilized in more than ten thousands research articles over the last ten years. Although there are excellent review and comparison studies for most of the research that utilizes Twitter, there are limited efforts to map this research terrain as a whole. Here we present an effort to map the current research topics in Twitter focusing on three major areas: the structure and properties of the social graph, sentiment analysis and threats such as spam, bots, fake news and hate speech. We also present Twitter's basic data model and best practices for sampling and data access. This survey also lays the ground of computational techniques used in these areas such as Graph Sampling, Natural Language Processing and Machine Learning. Along with existing reviews and comparison studies, we also discuss the key findings and the state of the art in these methods. Overall, we hope that this survey will help researchers create a clear conceptual model of Twitter and act as a guide to expand further the topics presented.

1. Introduction

Twitter is one of the most vital and vibrant Online Social Networks (OSN) today. It is commonly ranked as one of the most popular OSNs by having 650 million registered users, although practically it is the third most popular after Instagram and Facebook. This is because the top ranked, Google+ which was shut down on 8th of October 2018 (Abner, 2018), used to consider as active all users having a Google account (Wong & Solon, 2018). As reported by Alexa, Twitter is currently ranked as the 49th most popular website of the world (Alexa Internet, Inc., 2018). Twitter has 330 million monthly active users, 152 million daily active users and it accommodates 500 million tweets per day (LiveStats, 2018; Omnicore, 2018). All this information has established Twitter as a very important online social network for user interaction and information dissemination. It is estimated that the average daily engagement of American users in social networks is more than 3 h (Marketingcharts, 2013). Also, 45% of Americans between the age of 18 and 24 years old are Twitter users (Smith & Anderson, 2018).

Twitter has a simple data delivery model which is implemented in a very efficient and scalable infrastructure (Hashemi, 2018). Moreover, Twitter stands out from other OSNs from the fact that, although it has a typical OSN structure (users connected to users), it is mainly used for news dissemination (Kantrowitz, 2018; Kwak et al., 2010). This is due to the fact that accounts that represent public and private institutions, news agencies, public figures, music bands, political parties and other collectives of various nature, flourish on Twitter. These entities, along with accounts that represent individual users, make Twitter a very interesting research object in numerous areas like computer, social, urban and art sciences.

Google Scholar lists 27,000 research articles that include the word 'Twitter' on their title. In Fig. 1 we present a breakdown of these articles from 2006 (when Twitter was founded) until 2020, compared to the relevant number of articles that contain the word 'Facebook'. Yet, in contrast to Facebook (Wilson et al., 2012), very little effort has been

☆ This document is the results of the research project funded by the European Commission, project CONCORDIA, with grant number 830927 (EUROPEAN COMMISSION) Directorate-General Communications Networks, Content and Technology.

* Corresponding author.
E-mail addresses: despoina@ics.forth.gr (D. Antonakaki), fragopou@ics.forth.gr (P. Fragopoulou), sotiris@ics.forth.gr (S. Ioannidis).

https://doi.org/10.1016/j.eswa.2020.114006
Received 6 February 2020; Received in revised form 9 July 2020; Accepted 11 September 2020
Available online 12 September 2020

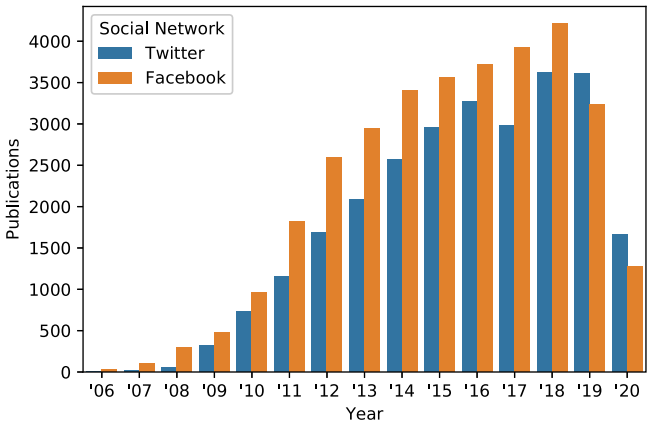

Fig. 1. Number of scientific publications containing the word 'Twitter' and 'Facebook' on their title from 2006 until 2020, as stated in Google Scholar. Although Facebook contains six times more users, in research, Twitter is only slightly less popular. The small decrease over the last years is not a sign of decreasing interest, but it is due to a batch effect: recent submitted studies have not yet been published.

made to map this new area, pinpoint the major methods, clarify the objectives and list the most important existing work. Here we attempt to present a survey of the major research themes and strategies for data analysis on Twitter. This review does not cover the numerous data formats and programming libraries that are available for this purpose. For a review of this area we recommend the comprehensive work of Batrinca and Treleaven (2015). See also Freelon (2018), for a comprehensive list of available online tools and resources for social media data collection.

1.1. The Twitter data model and basic terminology

A post in Twitter, is a short message, called 'tweet', with no more than 280 characters which used to be 140 until November 2018 (Rosen & Ihara, 2018). A tweet may contain images, URLs and videos. All accounts are by default public, meaning that any user can read the tweets of this account. A user can select to 'follow' any other public user. The 'timeline' of a user contains the temporal updates of the tweets of the users that she follows. Therefore, any user has a set of 'followers' (users receiving the tweets that this user sends) and 'followings' (users whose tweets appear on this user's timeline).

A user can configure the privacy settings of her profile to render her tweets as public or protected. Protected tweets are only visible to the users that are pre-approved by the original sender. Additionally, users on Twitter, can create lists with other users' accounts, or subscribe to a list created by another user. This will result in a more focused view of the timeline, since it will contain only tweets originating from users belonging to this list.

Apart from text, images, videos and URLs, a tweet may contain hashtags and user mentions. Tables 1 and 2 contains six of the most prominent features of this data model, along with the major research publications that explore them. These features are hashtags (HT), trends, retweets (RTs), mentions, replies and URLs. *Metrics* refers to the prevalence and the basic statistics of the feature. *Success* refers to the studies that assess whether or not this feature is a 'success' indicator of real world entities (i.e. scholar articles, political campaigns etc.). *Influence* refers to the studies that measure the contribution of this feature on user's influence. *Retweets* contains the studies that measure if the use of a feature increases the chances of being retweeted. *Has Graph* contains the studies that construct and analyze graphs with this feature as nodes. *Spam* contains the studies that measure the prevalence of this feature on tweets containing spam URLs. *Bots* contains the studies that assess whether this feature has been exploited by bots.

Finally, users can 'like' (or 'favorite') a tweet, although this feature is gradually phased out by the service, as reported in Telegraph (2018).

1.1.1. Hashtags

A hashtag is a word that is precedented with a hash (#) character, (i.e. '#funny'). These words are indexed separately and users can query the platform in order to find tweets with specific hashtags. Hashtags have evolved to a social phenomenon and their use has been adopted by several online and non-online media as a simple method to signify, idealize and conceptualize a single word (or phrase) in a short message. The general action of assigning hashtags to events, places or people has been described as 'social tagging', as shown in Huang et al. (2010) and is a vital part of Twitter and of microblogging in general. Metrics that are applied in hashtags are frequency, specificity, consistency and stability (Patel-Schneider et al., 2010). Frequency measures the amount of users and messages that contain it. Specificity measures the semantic relationship between the hashtag as a word and the context for which is used for. Consistency is the level of the hashtag's spread over different communities, as a referrer to a specific concept. Finally, stability is how the hashtag maintains both its frequency and its thematic content over time.

1.1.2. Trends

Popular hashtags and common search terms are listed separately, as popular 'trends'. In the literature these are also referred to as 'topics', 'popular trends' or 'trending topics'. Trends are different per geographic region and the topics that users view are determined by their location and the interests of the users that they follow. The study of Twitter's trends gives valuable information of the importance, duration and impact of real world events. For example, an interesting question is, if Twitter is a fresh content generator, or if it simply reproduces content from external sources. Studies show that Twitter acts as a content aggregator, driving specific trends to popularity (Asur et al., 2011). Moreover, there is a qualitative difference between trends that are emerging from user's activity and traditional headlines that are posted by mainstream media. Specifically, the events that appear first as Twitter trends, are usually captured by individual users (accidents, demonstrations, happenings, etc.), in contrast to political events that are covered mainly by professional reporters. Finally, 1 out of 5 users tweet about a certain trend and 15% participate in more than 10 topics, within a period of four months (Kwak et al., 2010).

A distinct area of research in Twitter is the semantic analysis of trends and hashtags. The purpose of these studies is to locate semantic relationships between trends and build a trend similarity graph (Wang et al., 2014). This graph can help pinpoint emerging topics (Naaman et al., 2011), categorize users into groups of interests (Abel et al., 2011a) and uncover hidden relationships between seemingly unrelated topics (Cataldi et al., 2010a).

1.1.3. Hashtags and topic recommendation systems

In topic analysis studies, hashtags are good predictors of the thematic subject of a tweet (Abel et al., 2011b), or event detectors (Adedoyin-Olowe et al., 2016).

This in turn, makes hashtags valuable in recommendation systems, trying to assist users to assign appropriate hashtags to tweets (Efron, 2010). Similarly, topic recommendation systems attempt to extract topics that are subjective to users' interests and friendships, while being timely and accurate (Cataldi et al., 2010b).

1.1.4. Retweets, mentions, replies and URLs

Users can 'retweet' or else re-post a tweet from another user. Users can also refer explicitly to a specific user by adding a 'mention' in a tweet, which is the character '@' followed by a username (i.e. '@jack'). On both events, the referred (retweeted or mentioned) user gets notified by the service. The number of retweets is commonly associated with the content-value of a specific tweet, whereas the number of mentions is associated with the name-value (or else fame) of the user (Cha et al., 2010).

Table 1
Some of the major studies that examine the basic features of Twitter. HTs = Hastags, RTs = Retweets. On the last two categories of Spam and Bots, underlined cells indicate that the respective studies reached negative conclusions (it is not exploited). Blanks (–) indicate that no study was found that measures the property of this feature. This table is continued on Table 2.

	HTs	Trends	RTs
Metrics	Patel-Schneider et al. (2010)	Kwak et al. (2010)	Lerman and Ghosh (2010)
Success	Asur et al. (2011)	Abel et al. (2011b)	Asur and Huberman (2010)
Influence	Suh et al. (2010)	–	Cha et al. (2010)
Retweets	Suh et al. (2010)	Suh et al. (2010)	–
Has Graph	Ferrara et al. (2016)	Cataldi et al. (2010a)	Bakshy et al. (2011)
Spam	Benevenuto et al. (2010)	Martinez-Romo and Araujo (2013)	Grier et al. (2010)
Bots	Ferrara et al. (2016)	–	Stella et al. (2018)

Table 2
Continued from Table 1. This part presents Mentions, Replies and URLs.

	Mentions	Replies	URLs
Metrics	Midha (2014)	Duncan (2010)	Wu et al. (2011) Bakshy et al. (2011)
Success	Thelwall et al. (2013) Tumasjan et al. (2011)	–	Dong et al. (2010)
Influence	Cha et al. (2010)	Ye and Wu (2010)	Eysenbach (2011)
Retweets	–	–	Suh et al. (2010)
Has Graph	Conover et al. (2011)	Bliss et al. (2012) and Nishi et al. (2016)	Bakshy et al. (2011)
Spam	Grier et al. (2010)	–	Ghosh et al. (2012)
Bots	Stella et al. (2018)	Ferrara et al. (2016)	Chu, Gianvecchio et al. (2012)

Although retweeting is one of the most known features of Twitter, in 2010 it was estimated that only 6% of tweets got at least one retweet (Duncan, 2010). Therefore, a common line of research is to locate the features that drive a tweet to be more retweeted (Boyd et al., 2010; Suh et al., 2010). One of these studies showed that tweets getting more retweeted, have similar textual and thematic content (Hong et al., 2011). Specifically, tweets with general thematic content (i.e. Christmas), or bad news are more likely to be re-tweeted (Naveed et al., 2011). Also, Suh et al. (2010) showed that URLs, hashtags, number of followers and followings affect positively the number of retweets, whereas the number of past tweets does not have any effect. In another study (Kupavskii et al., 2012), it was shown that the position of a user in the social graph (assessed by the PageRank metric) is also a crucial factor.

In cases where a tweet contains a URL promoting a future event, the ratio of the retweets before and after the event is a good predictor of its 'success'. For example Asur and Huberman (2010) used these metrics to predict the success of movies right after their release. in Eysenbach (2011) the author derived accurate predictions of the citations of a scientific paper, based on the number of tweets containing URLs to the online versions of this paper.

The total number of mentions that users receive is associated with the influence of their profile as a whole, and not with the impact of their individual tweets. For this reason, user mentions are commonly used to measure the 'success' of a user, as opposed to events. Examples that are employing the number of mentions, in order to measure the impact of a user profile, is on scholarly publications (Shuai et al., 2012; Thelwall et al., 2013) and election campaigns (Hong & Nadler, 2012; Tumasjan et al., 2011). The number of mentions is also useful for assessing the success of a paid advertising campaign on Twitter. This is because mentions require active engagement, in contrast to simple views. Twitter estimates that approximately 80% of its users have mentioned a brand at least once (Midha, 2014).

A user can reply to a tweet. Duncan (2010) estimated that 23% of tweets got at least one reply. Replies generate a typical thread, commonly seen in online forums. The reply-network is a graph, where nodes are users and edges represent reply events, in a defined period of time. This network is believed to represent user similarities in a higher level than that of the typical users–followers network, as studied by Bliss et al. (2012). Another type of reply-network is the reply-cascade tree, which simply represents the discussion thread initiated by a single tweet, as shown in Nishi et al. (2016). The shape of this tree is highly dependent on the *indegree* (number of followers) of the root node (Nishi et al., 2016). The number of replies that a user receives is also a metric of influence (Ye & Wu, 2010). Both reply-networks and mention-networks can be used to measure the information diffusion pattern for certain events (or hashtags). Information diffusion measures the temporal variation of the network as information travels (diffuses) through its edges. For example we might be interested in separating 'normal' diffusion events (i.e. a user's reply) from undesirable (i.e. spam, hate) (Foroozani & Ebrahimi, 2019).

Finally a common feature in tweets is the contained URLs. It is estimated that a percentage between 10% (Bakshy et al., 2011) and 20% (Suh et al., 2010) of tweets contain URLs. Given the hundreds of millions of daily posted tweets, it is evident that an active area of research is about the evaluation of Twitter as a general purpose web search engine (Teevan et al., 2011). Indeed, in a study of Dong et al. (2010) it was found that URLs posted in Twitter were more relevant and more recent, compared to results returned from common search engines. Also URLs have a big variety of life span, depending on the category of the poster. Wu et al. (2011) showed that URLs posted by media organizations are short-lived, whereas URLs from bloggers have a longer life span, especially if they link to music or video content. The same study also concluded that 50% of URLs posted in Twitter are 'generated' (or else initially posted) from a very low number of 'elite' users. In Section 4.1 we survey another crucial area of research, which is the presence of spam and other malicious URLs in Twitter.

1.2. Getting data from Twitter

Since its beginning, Twitter has made available an API (Application Programming Interface) for accessing not only data but also most of the service's functionality. Initially, Twitter used to have a very open policy regarding data access (Mersch, 2018). Taking a look at early papers regarding Twitter from 2010 (Kwak et al., 2010), we notice that it was possible to collect the entire social graph of Twitter in a period of 2 months, by using only 20 workers. Also it used to 'whitelist' IPs with unlimited access for research purposes (Benevenuto et al., 2010; Gabielkov & Legout, 2012). Fearing that third-party services could misuse the API and build applications that could essentially mimic its main functionality, Twitter started in 2012 enforcing more strict rate limits as reported by Twitter official blog (2018). API requests, should be authenticated through OAuth2 and are monitored on a per 15 min window (Twitter official API documentation, 2018). In this window, API requests are limited according to their type. For example, in order to request the timeline of a user, a client can perform 900 requests per allowed window. Each request can fetch up to 200 tweets and clients can only fetch up to 3200 of a user's most recent tweets.

Twitter has made available paid data access plans that have more relaxed limits. This is in concordance with Twitter's almost constant effort to monetize its service (Titcomb, 2018).

According to Twitter (2020), since November 2017 the API provides free search access to its data published in the past 7 days. 'Premium' and 'Enterprise' are paid API access models that provide access to tweets from the last 30 days or from as early as 2006 respectively. The

Table 3
A brief timeline of the major policy changes of accessing Twitter data. The examples column contains studies with sections describing how the authors employed the relative data access method. Tromble et al. (2017) discusses the effects in scientific inference of the fact that free and premium API return different sets of data.

Period	Policy	Examples
2006–2010	Whitelisting, Relaxed limits	Kwak et al. (2010) and Benevenuto et al. (2010)
2010–2012	OAuth2	McCreadie et al. (2012) and Wang (2010)
2012–2017	Strict API limits	Antonakaki et al. (2014) and Borra and Rieder (2014)
2017–Today	Premium/Enterprise APIs, No public data	Mazza et al. (2019) and Tromble et al. (2017)

Enterprise API access requires an application and the pricing for the Premium depends on the query. The Premium API access allows for 500 tweets per request and 60 requests per minute. Prices range from $149 for up to 500 requests per month till up to $2499 for 10,000 requests per month. In 2019 13.5% of Twitter's revenue originated from data licensing and other sources making in total $0.5 billion (Investopedia, 2020). In Table 3 we show a brief timeline of the major changes on the data access policies in Twitter.

Since the free access policy, the prices and the different access options are changing over time, it is difficult to estimate the percentage of publications that have paid for data access. In a 2017 review that examined 108 health related studies that used data from Twitter (Sinnenberg et al., 2017), only 44 (41%) used a free service. Besides the price, another concern is that the free API does not return the same results as the paid options. This has alarmed researchers since the discrepancies between free and paid options is high enough to produce different results in content analysis (Tromble et al., 2017).

To overcome these limitations, researchers usually utilize multiple applications, created from multiple fake accounts, risking violating Twitter's Terms of Service and getting their applications suspended. Besides using Twitter's API that has the limitations presented above, another option is to build scripts that imitate the actions of a browser when visiting the twitter.com page. This technique is called crawling or scraping which employs advanced web access methods and is itself an area of research (Borra & Rieder, 2014). An example of a crawler (or scraper) is TwAwler (Pratikakis, 2018) that can crawl the complete set of tweets, following relationships and other meta-data of an entire community, as big as the Greek (about 330 thousands), by using a single authenticated user and a usual desktop PC. Another crawler that focuses on medium size communities is TwitterEcho (Bošnjak et al., 2012). Other similar implementations that do not rely on Twitter's API and can access unlimited historic data are GetOldTweets3 (Mottl, 2020) and the one published from Hernandez-Suarez et al. (2018). It should be noted that these methods are for accessing raw data from Twitter. A third option is to use one of the many integrated frameworks that allow both data access, data exploration, filtering and analysis. An overview of more than 20 tools of this kind is available at Ahmed (2020). From these tools, 7 are free and 3 are low cost (less than US$100) and are systematically reviewed in Yu and Muñoz-Justicia (2020).

Researchers should be aware that releasing Twitter data violates the Terms of Services.[1] As an effect, large and well-studied Twitter datasets, like the Edinburgh Twitter Corpus (Petrović et al., 2010) and the SNAP dataset (Yang & Leskovec, 2011) are now not publicly available. The unavailability of public Twitter data has severe effects on the measurement of the reproducibility of current research. Also the absence of 'gold standards' has stripped the community of the ability to perform comparison studies. Today, there are two suboptimal approaches to circumvent these limitations. The first is to obtain releases of anonymized and heavily processed data.[2] The second approach is to use releases that include only the unique IDs of tweets and let researchers obtain the rest of the data with their own means. Examples of the latter are, the TREC 2011 Microblog Track (McCreadie et al., 2012), the SemEval Twitter datasets (Nakov et al., 2016) and the Stanford Twitter Sentiment Data (Go et al., 2009).

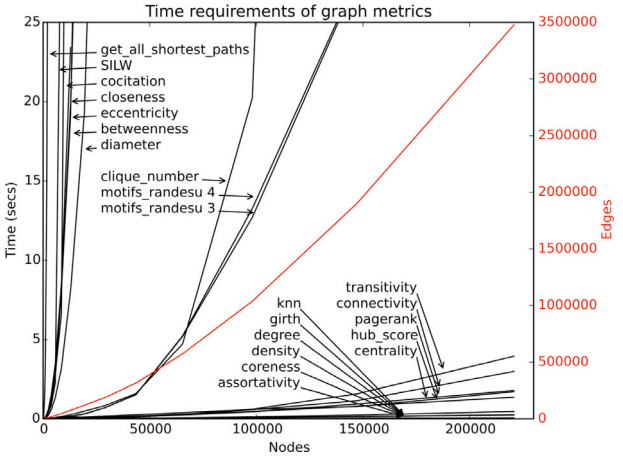

Fig. 2. Time required for estimating graph metrics in a social graph containing 13.2 million nodes and 8.3 billion edges. SILW = similarity inverse log weighted, KNN = K-Nearest Neighbord, Motifs RAND-ESU is a method for locating small motifs of size 3 and 4 (Wernicke & Rasche, 2006), hub_score is the Kleinberg's hub score (Kleinberg et al., 1999).

Twitter's API returns data in JSON format, with a relatively complex structure. Therefore, researchers show a preference towards NoSQL databases that natively support JSON structured data, like MongoDB, instead of performing complex conversions required for conventional relational databases.

1.3. Sampling the social graph

Before conducting any study in social networks, researchers have to make a crucial choice: the sampling technique. The size of the social graph of Twitter is in the range of hundreds of millions nodes and hundreds of billion edges. This size makes sampling a necessary prerequisite step, mainly due to (1) API access limitations and (2) extreme computational resources for storage and processing large networks. Apart from some basic graph measurements (for example average node degree), some of the most informative measurements are in the computational order higher of O(N). This renders these calculations practically impossible on the complete social graph of modern social networks. To demonstrate this we downloaded a subset of Twitter's social graph containing 13.2 million users and 8.3 billion edges with the Random Walk technique (Leskovec & Faloutsos, 2006). This technique is presented in detail later. The sampling took place in January 2015. Subsequently we used the igraph library (Csardi & Nepusz, 2006) in order to measure the time required to assess a variety of graph properties in a single CPU computer with 8 GB of RAM. In Fig. 2 we show that there is a family of important graph properties like diameter and betweenness centrality whose time complexity is exponential to the number of graph nodes.

To overcome these computational constraints, an efficient sampling technique has to be selected prior to any analysis. We can make two

[1] Currently it allows the public release of up to 50,000 tweets per day, per user. See term I.F.2.a on Developer Agreement and Policy: https://bit.ly/3iHiZNg.

[2] A list of Twitter datasets and related resources: https://github.com/shaypal5/awesome-twitter-data.

large distinctions of sampling methods. The first category disregards the user's activity and focuses solely on the network attributes. The second category also takes into account the user's activity.

Leskovec and Faloutsos (2006) applied 10 different sampling techniques to various social networks and measured how well each technique captured the properties of the networks. From all sampling techniques Random Walk and Forest Fire exhibit the best performance. Random Walk is the sampling method where a node is selected by random and is used as a starting point for a random walk in the graph. Forest Fire is the method where we randomly select a node and then we simulate a fire by randomly burning adjacent edges and nodes. The nodes and edges that are not burned are finally selected. Leskovec and Faloutsos (2006) also estimated that a good sampling size should preserve at least 15% of the original size, in order to match the most significant graph properties, such as the average in and out degree.

Although this was an excellent analysis, it has several practical problems when it comes to modern social networks. The most important is that the authors performed their experiments on large social graphs of their time (2006). At that time Twitter did not exist and Facebook had approximately 50 million users. A sampling size of 15% is still prohibitive for modern social networks, in terms of computational resources. Another limitation is that they do not take into account other valuable information that might make a node worth of sampling. This is the user's activity and user's influence.

In a later study of 2010 (Choudhury et al., 2010), researchers tested sampling methods that combined common sampling techniques, with information regarding user's activity and location. They also used different measurement methods that took into account the ability of the sample to capture 'diffusion events'. A diffusion event is the spread of a trend, a URL or a retweet. They concluded that the sampling method that exhibited the lowest distortion from the original graph is the Forest Fire, combined with activity information. They also estimated that an optimal sampling size is approximately 30%. Nevertheless, the validation of this technique on Twitter is an open question. Perhaps the largest sampling experiment that has been performed on Twitter is from Gabielkov et al. (2014a), which sampled the complete social graph as of 2012. This study concluded that common sampling techniques like Breadth-first search, Random Walk and an alleged unbiased sampling technique, suggested by Wang et al. (2011), are all biased towards high degree nodes. Therefore the optimal sampling technique and sampling size is to a certain extent, an open question. Since today it is practically impossible to acquire an adequate subset of Twitter's social graph, researchers choose to focus in special subgroups such as the verified users (Paul et al., 2019) or the top celebrities (Motamedi et al., 2020).

An orthogonal question for social networks is what is the best method to generate artificial graphs, with properties similar to real social graphs. in Leskovec, Lang et al. (2008), the authors explored 70 sparse real networks and found that a generative model built with the 'forest fire' technique burning process, can produce a graph with similar community structure to the real ones.

1.4. Generating time snapshots

Twitter does not reveal the creation time of the edges (followings) and therefore it is practically impossible to generate a precise snapshot of the social graph, for a given time. This policy, along with the general restrictions of Twitter's API, has generated criticism, since valuable historic data are extremely difficult to obtain Batrinca and Treleaven (2015). Nevertheless, Twitter provides the exact time a node was created (user registration) and also the lists of followers and friends of a user, ordered according to the following creation time. These two pieces of information can be combined to produce a lower bound estimation of the following creation time (Meeder et al., 2011). The accuracy of this heuristic depends on the number of friends or followers of a user. For users with more than 5000 followers, the link creation time is estimated within an accuracy level of several minutes.

Gabielkov et al. (2014b) performed a temporal analysis of Twitter's macrostructure, by using only the user creation time. Also, Antonakaki et al. (2018) used this exact heuristic to measure the evolution of Twitter's average node degree from 2006 until 2015.

1.5. Taxonomy of Twitter studies

In the following chapters we analyze three of the most important research areas on Twitter. These are the Social Graph, Sentiment Analysis and Threat Detection. Before moving on, we believe that it is important to put these three areas in a unified schema, along with the already presented 'Basic' theory of Twitter. This schema is presented in Fig. 3 and places *Twitter* in a central position where four main branches depart. Under these branches we place the basic concepts for each area. Of course this schema is largely incomplete but it contains a rough mental model of a very complex and dynamic environment. We hope that it will help the community to refine it and to expand it with future studies.

2. The social graph of Twitter

The social graph of an OSN is defined as the graph on which vertices (or nodes) represent users and edges (or links) represent following relationships. Or else, if user A follows B, this is represented on the graph with a directed edge from node A to node B. Twitter's social graph is directed, which is not always the case in OSNs. For example, in Facebook, a 'friendship' is established after a mutual agreement between two users, therefore the formed social graph is undirected.

The social graph has been the center of attention in many research areas. Various properties of this graph are indicative of the nature of the social network and portray how users perceive the platform and interact with other users. It can also provide insight on the temporal dynamic of the platform and the well-being of the service.

Here, we separate studies on the social graph on Twitter in two major categories. The first category studies the social graph at the node level, trying to infer methods that measure the influence, popularity and the social impact of individual users. The second category studies the social graph as a whole, trying to understand the structure and the high-level dynamics of the network.

2.1. The social graph at the node level

There are two areas of studies that focus on the node level of the social graph. The first contains metrics that try to assess various aspects of the user's presence on the network. Although there exists an ambiguity regarding what is 'presence' and how to assess it, a review from 2016 (Riquelme & González-Cantergiani, 2016) suggested that there are three distinct attributes that can be measured. These are *activity* which measures how frequent a user interacts with the network, the second is popularity which measures how well a user is recognized from other users and the third is *influence* which measures how a user's actions influence the actions of other users. The attribute that has attracted the higher interest from the community is the influence, which we will present in detail below. The second area of studies focusing on the node level of the social graph is the phenomenon of *homophily*.

2.1.1. User influence

In the early stages of Twitter, reaching a high number of followers was considered a strong indication of a user's influence. In one of the most cited papers regarding Twitter, it was shown that the number of followers (*indegree*) is not related to the number of retweets and mentions that a user receives (Cha et al., 2010). Indeed, events that require active user engagement (like retweets and mentions) are better estimators of a user's influence compared to passive followings. Moreover, a tweet from a user with low number of followers, can reach

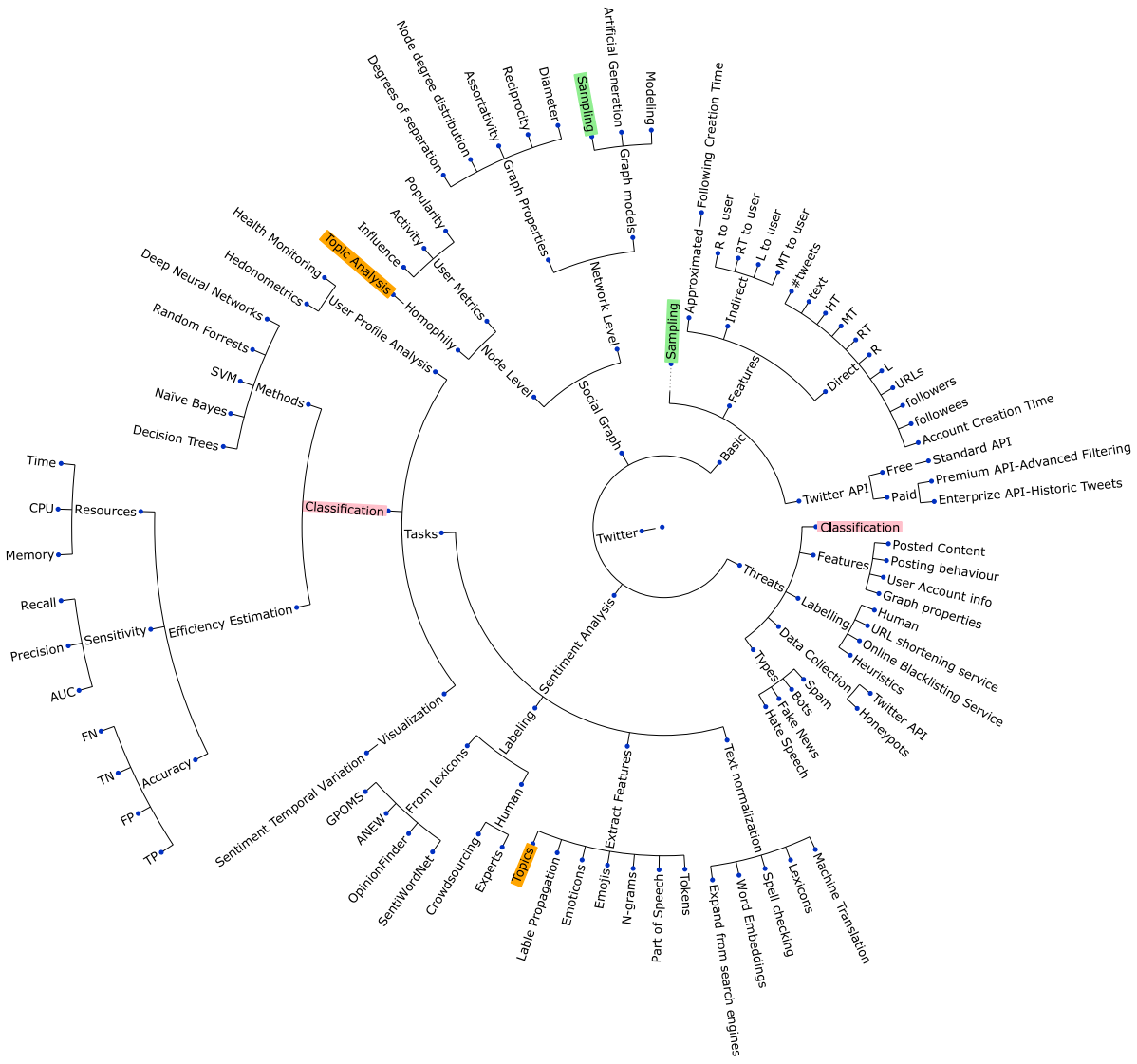

Fig. 3. The taxonomy of the most important concepts in research involved in Twitter. The four main branches are 'Basic' containing data access methods and Twitter Features (i.e. Hashtags), 'Sentiment Analysis', 'Social Graph' and 'Threats'. Although the schema has a tree structure (shown in circular layout), some nodes (concepts) exist in more than one area. These are the 'Sampling' node which exist under the 'Basic' and the 'Social Graph' branch, the 'Classification' node that lies under the 'Sentiment Analysis' and 'Threats' branch, and the 'Topic Analysis' node that lies under the 'Sentiment Analysis' and the 'Social Graph' branch. HT = Hashtags, MT = Mentions, RT = Retweets, R = Replies, L = Likes.

orders of magnitude higher number of users through retweets (Lerman & Ghosh, 2010) (the number of users that a tweet finally reaches is also called *impressions*).

These findings sparked an interest for metrics that can give more insightful indications of a user's influence by also taking into account the user's position in the social network (or else user's topology). In a very thorough review, Riquelme and González-Cantergiani (2016) presented the characteristics, kinds of metrics used and the time complexity of 53 different influence metrics. Most of these belonged to the PageRank family of methods. Here, we discuss two of the most known metrics for influence, which are PageRank and Betweenness Centrality.

PageRank Regarding topology, the most widely known algorithm that measures a user's influence in a social network, is PageRank (Brin & Page, 1998). PageRank was initially designed by Google to measure the relevance of a website on the Internet, regarding a search term. To assess the value of PageRank for all nodes, we initially assign small constant (or random) values to all nodes of the graph. Then, for each node, we re-assign this value to the weighted sum of the PageRank of all nodes that link to this node. We iterate this step until the assigned values converge. Once it converges, a node with high PageRank value means, in general, that it has other nodes with high PageRank values linking to it. The application of the same algorithm in social graphs indicates the influence of a user. Studies that have applied PageRank on Twitter include: Kwak et al. (2010), Said et al. (2019) and Weng et al. (2010). Additionally, in Priyanta et al. (2019) the authors suggest that the PageRank of a node is directly proportional to the closeness centrality and the in-degree of this node.

Variations of PageRank exist that are more tailored to the context of Twitter, like the Influence-Passivity Algorithm (Romero et al., 2011). Also the Hirsh index (Hirsch, 2005) resembles PageRank, since it is an influence measurement algorithm that also originates from a different concept (measurement of scientific impact) and can be adapted to social networks.

It is interesting that these studies have done limited work in comparing PageRank with other methods and they limit their analysis on presenting the top popular users on Twitter according to this metric. This might be part of the wider issue, that there is not any gold standard or a widely accepted methodology for assessing the accuracy of user influence methods. PageRank (and its variations) is utilizing solely the

network structure, in order to assess a user's popularity. Nevertheless, we have seen that topic modeling is also important for measuring the influence of a user, in respect to a certain topic. Consequently, the combination of these methods (PageRank and Topic Modeling) is a far more powerful approach for measuring user's influence (Haveliwala & H., 2002).

Betweenness centrality Betweenness centrality is the ratio of all possible shortest paths that pass from a certain node. A node with betweenness centrality equal to 1.0 means that it exists in every shortest path, between any two random nodes of the graph, indicating a maximum influence. This metric, although computationally expensive, can be fairly approximated by sampling a small number of nodes in artificial networks (Bader et al., 2007). Despite this, it is very sensitive to noise since an extra node can alter significantly its value. A variation of betweenness centrality that is computationally lighter and more robust to noise is the K-Betweenness Centrality, which takes into account nodes that lie at most k edges away (Madduri et al., 2009). A specially designed system for measuring Betweenness centrality on Twitter is GraphCt (Ediger et al., 2010), which employees this metric in order to locate key users for a given topic.

Betweenness centrality belongs to a large collection of measures that try to assess the importance of a node in the network, with information solely from the network topology. Other interesting measures in this family are the closeness centrality which is the average shortest path length with all other nodes (Priyanta et al., 2019), the eigenvector centrality which is a predecessor of PageRank (Howlader & Sudeep, 2016; Maharani et al., 2014; Said et al., 2019) and Katz centrality. Katz centrality resembles PageRank with the fundamental difference that it takes into consideration all nodes of the graph (i.e. not only adjacent nodes), with a weight that is exponentially reduced according to distance (Hanneman & Riddle, 2005). An interesting variation of Katz centrality has been used to assess user influence, by also taking into account the temporal flow of information in the network (Laflin et al., 2013).

As it has been shown by Weitzel et al. (2012), there is a small correlation ($r^2 = 0.5$) between Betweenness Centrality and PageRank. The choice between these metrics depends on what we want to focus on. PageRank focuses on the quality (links to influential users) of a node and as an effect it is suitable for identifying nodes with local influence. Centrality focuses solely on the location of the node in the graph and is suitable for measuring global influence. PageRank has far more variations than Betweenness centrality and can be easily adapted in order to include other types of network information. Betweenness Centrality is sensitive to inclusion (or deletion) of very popular nodes (celebrities). Yet, Betweenness Centrality is relatively stable over time, in contrast to PageRank which is sensitive to small network alterations and it should be computed much more frequently (Riquelme & González-Cantergiani, 2016).

Tweets, Retweets and Followers Another straightforward metric for measuring user's influence is the number of retweets and the numbers of followers. Kwak et al. (2010) demonstrated that the number of followers is highly correlated to PageRank, whereas there is low correlation between followers and retweets and between PageRank and retweets. Surprisingly, Lerman and Ghosh (2010) reached a contrasting conclusion that there is a strong linear correlation between number of followers and retweets. The same study also showed that a good predictor for the number of followers is the number of followings, demonstrating that diversity of information in tweets is often rewarded by more retweets. Nevertheless, studies have shown that there is a strong correlation between number of followers and the number of different users that they usually retweet (Chun et al., 2008), thus when retweeting, users exhibit a strong favoritism towards certain users. The depth of the retweet pattern of a posted URL can give information for the significance of the URL and the influence of the user who first posted it (Bakshy et al., 2011).

In general, there is a positive correlation between number of followers and number of tweets (the more the followers, the higher the activity). It is indicative that in Twitter, 10% of users have 10 or lower followers and rarely tweet. At the other end, it is easy to spot 'celebrities' by measuring the ratio of tweets versus followers. This is because the positive correlation between the number of tweets and followers stops for users that have more than 10,000 followers. For these users, the high number of followers is due to their social status, rather than the quantity of their tweets (Kwak et al., 2010).

2.1.2. Homophily

Homophily is the level at which people with common interests tend to associate in public environments, like social networks (McPherson et al., 2001). In order to measure this property we need a dataset enriched with user's activity (or interests), since the social graph itself does not convey this information. For this reason, researchers are using either the text from the messages, or they are utilizing meta-information of the social graph provided by the social network service. In the first case, a common line of work is to perform a Topic Modeling analysis, in order to extract the different topics present in a set of messages. One of the most common methods for Topic Modeling analysis is the Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Griffiths & Steyvers, 2004). After topic modeling, we can assess the degree on which a user is interested in a specific topic. Incidentally, this also measures the influence of this user on this topic. Finally, the correlation between the proximity of users in the social graph and the degree of shared interests gives an estimation of the homophily (Weng et al., 2010).

When utilizing meta-information from the social graph, homophily is measured according to the similarity of the time-zone, popularity (number of followers) (Kwak et al., 2010), or the similarity of the subgraph in the vicinity of a user's node, like in the PageRank algorithm (Brin & Page, 1998). Another source for information regarding a user's interests is the self placing of a user's followers into certain lists. The name of the list can give information of the primal identity of other users. By exploiting this information (Wu et al., 2011) revealed that there is a strong homophily, as expressed in retweets among celebrities, bloggers and media (in declining order of homophily).

On Table 4 we briefly present the meaning of influence, centrality and homophily along with the major studies that investigate these features.

2.2. The social graph as a whole

Graphs representing social networks have been a major research area long time before the era of OSNs. Freeman (2004) has presented a thorough historical analysis that starts from the beginning of the 20th century. In this section, we present the main efforts to measure properties of the complete social graph of Twitter. A summary is available in Table 5. Each measurement gives us intrinsic insights of the characteristics and dynamics of the network.

2.2.1. Degree of separations

Perhaps the most notorious property in social networks is the 'degree of separations'. This is the average number of 'hops' that we need, in order to traverse the social graph from any random user to any other random user. This property has become famous even from the early ages of social studies, in 1967, when it was discovered that two random individuals are no more far apart than 6 hops in the social graph that represents real life (no virtual acquaintances) (Milgram, 1967). This finding was publicized as 'six degrees of separation' or else the 'small-world' phenomenon. Due to the easy way in which relationships can be created in social networks, this number is expected to be smaller.

This property is usually assessed by measuring the average length of all shortest paths of the network. This was measured to 4.12 on Twitter (Kwak et al., 2010), 2.74 on Twitter's subset of verified users

Table 4
The most common user properties related to the social graph: influence, centrality and homophily.

Property	Meaning	Measurements
Influence	How important are users who follow me?	PageRank: Kwak et al. (2010), Said et al. (2019), Weng et al. (2010), Priyanta et al. (2019) Hirsch Index: Haveliwala and H. (2002) Relation, Interaction graph: Räbiger and Spiliopoulou (2015)
Centrality	How centrally am I placed on the social graph?	Betweenness Centrality: Ediger et al. (2010), Priyanta et al. (2019) Eigenvector Centrality: Said et al. (2019), Howlader and Sudeep (2016) Closeness centrality: Priyanta et al. (2019)
Homophily	How close am I to other users having the same interests as me?	Similar profile: Kwak et al. (2010) TwitterRank: Weng et al. (2010) Lists: Wu et al. (2011)

Table 5
The most common properties of the social graph of OSNs and their respective measurements on Twitter.

Name	Measurement	Study
Degree of separation	Average Node Distance (AVN) = 4.8 AVN for Verified users = 2.74	Kwak et al. (2010) Paul et al. (2019)
Distribution of node degree	In-degree is power law $\lambda = 1.35$ Out-degree is log-normal $\mu = 3.56$, $\sigma^2 = 2.87$	Myers et al. (2014)
Average node degree	Average followers: 557.1 Average followings: 294.1	Bakshy et al. (2011)
Assortativity	The number of my X is the same as the number of Y of the users that I follow: X = Friends, Y = Friends: 0.272 X = Followers, Y = Friends: 0.241 X = Friends, Y = Followers: −0.118 X = Followers, Y = Followers: −0.296	Myers et al. (2014)
Interest assortativity	Music: 0.737 Sports: 0.811 Cinema: 0.776	Buccafurri et al. (2016)
Reciprocity	Percentage of users that follow me back 78% Reciprocity for 10K-elite Twitter network Ranges 31%–42%	Kwak et al. (2010) Motamedi et al. (2020)

(Paul et al., 2019), and 4 on Facebook (Backstrom et al., 2012). Another property, the diameter, is the longest of all possible shortest paths and portrays the linear size of the network. Since, this property is sensitive to distant outliers of the network, the 90th percentile is more commonly used, called the 'effective diameter' (Leskovec et al., 2005), estimated in 2010 to be 4.8 (Kwak et al., 2010) for Twitter.

2.2.2. Distribution of node degree

Another important property of social networks is the node degree distribution. In directed graphs, like Twitter's OSN, the degree of a node is the sum of the *outdegree* and the *indegree* property. The *outdegree* is the number of edges with direction outward to the node, whereas *indegree* is the number of inward directed edges. The average node degree is a measurement of the density of the graph and characterizes the amount of user inter-connections in the network. The average degree has a significant meaning on the modeling of users' behavior, since it has been associated with Dunbar's Number theory, which states that humans can have a finite number of stable social interactions in the range of 100 to 200 (Bliss et al., 2012; Gonçalves et al., 2011).

Most importantly, if the log-distribution of the node degree follows a power law (Mislove et al., 2007), then the graph is a scale-free network (Milgram, 1967; Travers & Milgram, 1969). Scale-free networks take their name from their general property to have similar structure to parts of themselves (also called self-similarity). The exponent λ of the power law for most real-life, scale-free networks is a value in the range from 2 to 3. Kwak et al. (2010) measured the exponent of the power law for this distribution in Twitter to 2.276. A study of the same year which examined the topology of 54.3 million users (Sadikov & Martinez, 2009), found that both the outgoing and incoming degrees follow a power law, with exponents 1.95 and 2.13, respectively. Nevertheless, a study of 2010, with 41.7 million users (Kwak et al., 2010), concluded that Twitter deviates from other social networks and that the outgoing degree distribution is not a power law. The most recent study (Myers et al., 2014) with the largest sample size (175 million users) estimated that the *indegree* is best fitted by a power law, with $\lambda = 1.35$, whereas the *outdegree* is best fitted by a log-normal distribution, with $\mu = 3.56$ and $\sigma^2 = 2.87$. Given the plethora of contradicting findings, we can conclude that the elucidation of Twitter's degree distribution is an active research question. Nevertheless, all studies agree that Twitter follows a *partial* power law, if we restrict it to users with less than $\sim 10^5$ followers. This property also contributes to the 'small-world' phenomenon described before.

It is important to note that whether the average in and out degree is a power law or not, has an important practical consequence. The mean of a power-law distribution with exponent $\lambda < 2$ diverges, or else it is not strictly defined (Newman, 2005). If this is the case for Twitter, there is no point in assessing the average node degree, despite being a very simple and intuitive measurement. Nevertheless, as stated in a study of at 2011 (Bakshy et al., 2011), the average *indegree* (followers) of Twitter was measured at 557.1 and the average *outdegree*(friends) was 294.1. For the same year, the average node degree of the undirected social graph of Facebook was measured at 190 (Ugander et al., 2011).

2.2.3. Assortativity

Assortativity measures the correlation between properties of adjacent nodes. We can perceive assortativity as a method to measure the

average homophily of the network, if we focus only on node properties (and not on the contents of users' posts). The most common type of this measurement is the degree assortativity. It has been suggested (Johnson et al., 2010; Newman, 2002) that assortativity can distinguish social networks from other 'real life' networks. The rational is, that when a new edge (i.e. friendship) is formed in a social network, it tends to reach a node with similar attributes. This is not the case for other 'real life' networks, like biologic or distribution networks, which tend to reach maximum entropy and their assortativity index is negative (also called disassortative networks). Indeed, the assortativity of Facebook, at 2011, was 0.226 (Ugander et al., 2011). Kwak et al. (2010) argues that since Twitter's social graph is directed, measuring assortativity is not feasible. Nevertheless, Myers et al. (2014) measured all four possible degree assortativity indexes (in- and out-degree of the source combined with the in- and out-degree of the target) and found 2 assortative and 2 disassortative. As expected, if we limit our network to users with the same interests the assortativity can be tripled (Buccafurri et al., 2016).

Another way to bypass this limitation of measuring assortativity is to consider an undirected social graph, by taking only reciprocal connections (i.e. two users are following each other). This approach has been used to measure the assortativity of reply-networks by Bliss et al. (2012).

2.2.4. Reciprocity

The social graph can also indicate the primal purpose for which a user uses a social network. This can be measured with the 'reciprocity' value, which is the degree of which a user is followed by the users that she follows. A low reciprocity shows that the user is using the social network mainly as a source of information, whereas a high reciprocity shows that the social network is mainly used for communication among a users' peers. Reciprocity has been measured for several social networks like 68% for Flickr (Cha et al., 2009), 84% for Yahoo! (Kumar et al., 2006) and 78% for Twitter (Kwak et al., 2010). If we define as celebrities the users belonging to the top 10,000 positions ranked according to the number of followers, their reciprocity ranges from 31% to 42% (Motamedi et al., 2020). We will refer later to this group of users as the 10K-Elite.

2.3. Modeling the social graph

As with any model that describes natural entities, a well formulated model that generates artificial networks, with properties similar to those of real OSNs, is of extreme importance (Leskovec, Lang et al., 2008). The main design principle of a mathematical model for the evolution of modern OSNs is to be able to formally describe the behavior of users, in a way that the structure and properties of the network can be predicted over time (Kumar et al., 2006).

Some of the properties that have been observed in large OSNs are the 'rich get richer' property (Barabási, 1999), the 'small world phenomenon' (Kleinberg, 2000; Said et al., 2019) and the decreasing diameter (Leskovec et al., 2005). The 'rich get richer' property suggests that new nodes prefer to be connected with nodes with high degree. This is also known as the 'preferential attachment' process. The 'small world phenomenon' suggests that the average shortest path between two random nodes in the network is proportional to the logarithm of the network's nodes. The 'decreasing diameter' suggests that as the network grows, the diameter decreases over time, suggesting that the network 'shrinks' or becomes more dense.

Apart from the theoretical interest, these models can have significant impact on the design of practical tools. Examples are sampling techniques (Leskovec & Faloutsos, 2006) and following recommendation systems (Barbieri et al., 2014; Bliss et al., 2013; Seo et al., 2017). A concise model can help build effective defenses against attacks like bots, fake accounts (Ferrara et al., 2016) and spam campaigns (Benevenuto et al., 2010). Additionally, the area of community detection (Barbieri et al., 2013; Said et al., 2019) and measurement of users' influence (Bray, 2015; Morales et al., 2014) rely heavily on these models.

One of the latest and most widely accepted models is from Leskovec et al. (2007). This model challenged the existing belief that OSNs evolve with a constant average degree and a slowly growing diameter. In contrast, the authors suggested that, as new nodes are added to the graph, the average degree of modern OSNs is increasing, whereas the diameter is decreasing. This model has been extended by Kleineberg and Boguñá (2014), to incorporate the layer of the existing, yet unobserved, off-line social network. Although, this model has been validated in 70 small real-life social and information networks (Leskovec, Backstrom et al., 2008), efforts to validate it on larger social networks like Facebook (Backstrom et al., 2012) and Twitter (Antonakaki et al., 2018) are partial and inconclusive. The main reasons for this are the limits of Twitter's data access API and the large computational requirements of the validation methods. To put this in a perspective, the computational complexity of measuring the diameter property is in the order of $O(|V||E|)$ (as seen in Fig. 2), which in the case of Twitter, can reach the prohibitive amount of 10^{20} calculations for a sufficient sample size. Leskovec et al. (2007) also acknowledges the fact that modeling the diameter remains an open question.

The best effort to create a macrostructure of Twitter is from Gabielkov et al. (2014b). This study, identified the Largest Connected Component (LSS) of the social graph and grouped it as a single node. Subsequently, through breadth first search, they identified smaller components, targeting or being targeted by the LSS. Overall, this technique allowed not only the elucidation of Twitter's macrostructure, but also the exploration of the main patterns of information flow in the graph.

A more recent study of Motamedi et al. (2020), examined the 10K-Elite network of Twitter. This study provides the connectivity of strongly connected components and the basic connectivity features such as reciprocity, diameter and node degree distribution.

3. Sentiment analysis

One of the most promising methods for content analysis in social media is sentiment analysis (Giachanou & Crestani, 2016; Martínez-Cámara et al., 2014). 'Sentiment' usually is a variable that can take values like: 'Positive', 'Negative' and 'Neutral', or more specific values like 'Happy' and 'Angry'. Each variable can take a long range of values, allowing for multiple assignments of sentiment in a single word. This means that a word can have both positive and negative sentiment. Moreover, we can generate additional meta-features based on the sentiment values. These are 'subjectivity' and 'polarity'. Subjectivity is the ratio of 'positive' and 'negative' tweets to 'neutral' tweets. Polarity is the ratio of 'Positive' to 'negative' tweets. For example, sentiment analysis can measure the attitude of a group towards a specific issue or assess the 'positiveness' as a personality trait of a single user.

3.1. A common sentiment analysis pipeline

The usual methodology for sentiment analysis requires the preprocessing and extraction of lexical features from tweets (Kolchyna et al., 2015; Pak & Paroubek, 2010). Preprocessing steps can have significant effects on the performance of the model (Jianqiang & Xiaolin, 2017) and includes tokenization, expansion of abbreviations and removal of stop words and other elements without lexical value, like URLs and mentions.

This is an active area of research in NLP and is usually referred to as 'text normalization'. It is estimated that 15% of tweets have 50% or more Out Of Vocabulary (OOV) words (Han & Baldwin, 2011). Techniques include importing information from lexicons for abbreviation extension (Han et al., 2012), use of spell checking algorithms for word correction (Han & Baldwin, 2011), use of machine translation for sentence normalization (Kaufmann & Kalita, 2010) and use of Word

Table 6
The major steps of a general classification workflow for Twitter data with machine learning methods.

Step 1. Data collection	
Methods	Twitter API Borra and Rieder (2014), Hernandez-Suarez et al. (2018) and Pratikakis (2018)
	Open Datasets (Limited) Palachy (2018)
	Social Honeypots Lee et al. (2010, 2011) and Stringhini et al. (2010)
Step 2. Labeling	
Labels	Spam/Legit, Positive/Negative, Bot/Real, Rumor/True
Methods	Manual: Human Experts Mozetič et al. (2016)
	Crowdsourcing Finin et al. (2010): Amazon Mechanical Turk, Crowdflower
	Automatic: Spam: labeling from shortening services (t.co) Amleshwaram et al. (2013) and Gao et al. (2012) Sentiment: Emoticons Go et al. (2009) and Narr et al. (2012)
Step 3. Split data in 3 parts: Train, Test and validation	
Step 4. Feature extraction Aggarwal et al. (2012), Chu, Widjaja et al. (2012)	
Lexical	Tokens, Stems, POS, n-grams, stop words, emoticons
Content	URLs, Mentions, Hashtags, Topics, Date
Influence	Retweet, Reply, Like
Profile	Friends, Followers, Date of creation, Description, #Tweets, Date since last tweet
Step 5. Add knowledge from external resources	
Sentiment analysis	Sentiment lexicons and vocabularies Potts (2011)
Spam classification	Online Black-list services Amleshwaram et al. (2013), Grier et al. (2010) and Martinez-Romo and Araujo (2013)
Bot detection	Search engine results for the account name Flores and Kuzmanovic (2013)
Step 6. Machine learning	
Methods	Decision Trees Gao et al. (2012) and Martinez-Romo and Araujo (2013), Naive Bayes Wang (2010), SVM Benevenuto et al. (2010), Random Forests Lee et al. (2011) and Mccord and Chuah (2011), Deep Neural Networks Severyn and Moschitti (2015)
Implementation	Weka (Java) Hall et al. (2009), scikit-learn (python) Pedregosa et al. (2011)
Step 7. Estimate accuracy	
Basic metrics	True Positive, True Negative, False Positive, False Negative
Accuracy metrics	Precision, Recall, F1, Accuracy Benevenuto et al. (2010)
Sensitivity	Area Under the Curve (AUC), Confidence Intervals Naveed et al. (2011)
Efficiency	Time and resources needed for the complete workflow Grier et al. (2010)

Embeddings for measuring topic similarity (Fang et al., 2016). Also, it is known that short documents are not suitable for topic modeling (Hu et al., 2009). To overcome this, a very interesting approach is to perform a query in a search engine with the content of the tweet and augment it with the top results (Hu et al., 2012). Other methods are to add information from Twitter specific lexicon sets made with Machine Learning (Ghiassi & Lee, 2018) or to simply concatenate the tweets of the same user (Weng et al., 2010) and construct an author-based topic model.

Useful lexical features include word stems, Part of Speech (POS) tags (Derczynski et al., 2013) and n-grams. Fortunately, there are mature and efficient tools that perform these tasks, with minimal programming effort. Examples are NLTK (Bird et al., 2009) for Python and MALLET (McCallum, 2002) for Java.

Lexical features also include emoticons and emojis. Studies show that tweets with positive emoticons are four times more likely than tweets with negative (Speriosu et al., 2011), so researchers need to correct for this imbalance. Regarding emojis, the Unicode standard contains 2823 emojis and more than half of Instagram posts contain at least one (Dimson, 2018). Research has shown that most used emojis convey both positive and negative sentiment (Chen et al., 2018) and are valuable features for sentiment detection.

Other popular features are topics and entities. Topics represent clusters of common words that appear in a set of documents (Hong & Davison, 2010). A document can belong to multiple topics and topics do not necessarily have a 'real world' interpretation. The most common method for topic modeling is Latent Dirichlet Allocation (LDA) and a commonly used implementation for Twitter data is Twitter-LDA (Lo et al., 2017; Zhao et al., 2011). In contrast to topics, entities are notions with 'real world' meaning. The task of Named Entity Recognition (NER) is the extraction of a generic semantic identity for a word. For example 'Person' for 'Obama' and 'Place' for 'New York'. A popular NER tool is Stanford NER (Finkel et al., 2005), whereas T-NER (Ritter et al., 2011) and TwiNER (Li et al., 2012) are optimized tools for Twitter. Another methodology for topic detection is the Twitter Topic Fuzzy Fingerprints, used by Carvalho et al. (2017) and Rosa, Batista et al. (2014), Rosa, Carvalho et al. (2014). Also by importing the growth rate of word frequency we can significantly augment the task of topic detection (Choi & Park, 2019).

The manual labeling of sentiment in tweets is done through two possible methods. This first is through a panel of experts and the second is with crowdsourcing techniques. The crowdsourcing technique is the use of online platforms that allow anyone to manually label the tweets, usually with a small reward. Popular choices are CrowdFlower and the

Table 7
Comparison studies of various types of analysis in Twitter. FS=Feature Selection, TP=Text Preprocessing, ML=Machine Learning, DNN=Deep Neural Networks, S=Spam, SA=Sentiment Analysis, BT=Bot Detection.

Task	Study
FS for S	Herzallah et al. (2018)
ML for S	Herzallah et al. (2018) and Wu et al. (2018)
FS for SA	Prusa et al. (2015)
TP for SA	Jianqiang and Xiaolin (2017)
ML for SA	Gonçalves et al. (2013)
DNN for SA	Kim (2014)
ML for BT	Rodríguez-Ruiz et al. (2020)

Amazon Mechanical Turk (Finin et al., 2010). A very early analysis from Snow et al. (2008) argues that expert employment and crowdsourcing techniques produce both, equally qualitative results. A later study by Mozetič et al. (2016) discovered that the quality of manual labeling is more important than the choice of the classification method. A metric that is commonly used to measure the concordance of labeling among multiple workers is the Fleiss' kappa (Reiss, 1981). Interestingly, one method for locating and encouraging users to participate in a crowdsourced dataset labeling task is through Twitter bots (Alperin et al., 2017).

The result of all this pipeline is the construction of a feature rich dataset, that contains linguistic features and sentiments for the collected text from social media. This dataset usually is structured as a $T \times F$ vector space with T being the number of texts and F being the number of features. Alternatively, the extracted features can be modeled as graphs, by importing information from the social graph, via a method called 'label propagation' (Speriosu et al., 2011; Talukdar & Crammer, 2009).

This dataset can be used in a variety of methods. The first is to show the temporal variation of sentiment, over a course of a specific event. For example, we can visualize the variation of the sentiment of the public over the course of a political campaign, or a company event (Daniel et al., 2017). We can also quantify how specific actions or events altered the public sentiment. Another line of work is to build a machine learning classifier that predicts the sentiment of the public, based on the linguistic features. This can help to quickly assess the sentiment, based on linguistic features and find which linguistic features are more associated with sentiment. One of the most commonly used tools that provides most of the presented functionalities is Vader (Gilbert, 2014), which as reported by its authors outperforms even human annotators. For a review of available methods for automatic sentiment analysis see Gonçalves et al. (2013) and for a review of sentiment visualization techniques see Kucher et al. (2018). In Table 6 we present a typical classification workflow in Twitter, for a variety of classification tasks including sentiment analysis.

Also Table 7 contains a list of comparison studies for feature selection and text preprocessing techniques, machine learning methods and DNN architectures for sentiment analysis. Regarding features, as stated in Prusa et al. (2015), when only individual words (unigrams) are used as features, the best methods to measure the feature efficiency are Chi-Square and Mutual Information whereas the optimal number of features is 200. In the comparative study from Jianqiang and Xiaolin (2017), that examined different normalization techniques in a set of 5 open datasets, the authors showed that expanding acronyms and replacing negations significantly affects the task of classification. Also they demonstrated that Naive Bayes and Random Forest showed the highest sensitivity to the choice of normalization methods. In a comparison of 8 sentiment analysis lexicon based methods, Gonçalves et al. (2013) showed that SentiWordNet (Baccianella et al., 2010) exhibited the highest coverage (fractions of text with identified sentiment) and LIWC (Tausczik & Pennebaker, 2010) had the highest agreement (smaller deviations from the other lexicons).

3.2. Milestone studies, findings and notes

The first work on sentiment analysis in Twitter was performed by Go et al. (2009). This work used emoticons as sentiment indicators for labeling, used a train set of 1.6 million tweets and a test set of 300 manually labeled tweets. They extracted text features such as n-grams, bigrams and Part Of Speech tags and achieved a classification accuracy in the range of 82%. Many subsequent works used this study as a baseline, based on the fact that they also released the train dataset. Instead of emoticons, Kouloumpis et al. (2011), used hashtags for tweet labeling. Hashtags were manually labeled as positive (i.e. #success), negative (i.e. #fail) and neutral (i.e. #news). Liu et al. (2012) noticed that current models use either emoticons, or manually labeled tweets as sentiment labels for classification and suggested a hybrid system that imports information from both sources.

After these initial studies, we notice two parallel efforts in sentiment analysis. The first is to incorporate knowledge from external resources and the second is to measure the public opinion towards specific entities like persons, events and products.

Bollen et al. (2011) was the first study to employ an external lexicon, in order to label the sentiment features of tweets and associate their fluctuations with real events of 2008. This lexicon was the extended version of POMS (Profile of Mood States Pepe & Bollen, 2008), which contains 793 terms associated with 6 mood dimensions (Tension, Depression, Anger, Vigour, Fatigue and Confusion). Two studies from Saif et al. (2012a, 2012b) considered the use of entity extraction services (Rizzo & Troncy, 2011) like AlchemyAPI, OpenCalais and Zemanta and added entities in the feature set of tweets. Finally, Zhang et al. (2011) at the same year, used an opinion lexicon (Ding et al., 2008) tailored for product review analysis, to annotate the lexical features used for each entity of interest (they tested on Obama, Harry Potter, Tangled, iPad and Packers).

Examples of open existing dictionaries for NLP purposes and sentiment analysis are (for more see Potts (2011)):

1. SentiWordNet (Baccianella et al., 2010), a sentiment lexicon of 100,000 English words.
2. OpinionFinder, a subjectivity lexicon[3] containing 2.304 words, annotated as positive and 4.153 as negative.
3. Dictionary of English Stop words.[4]
4. The Affective Norms for English Words – ANEW – dataset. It contains emotional ratings for 1034 English words.
5. A Google-based Profile of Mood States (GPOMS) (Bollen et al., 2010). It assigns 6 emotion values (Calm, Alert, Sure, Vital, Kind and Happy) in any text, based on Google's n-gram collection.
6. The CMU Pronouncing Dictionary,[5] with pronunciation information for 134.000 English words.
7. The Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010), the 2015 version contains 6400 words, word stems, and emoticons.

Regarding sentiment measurement towards specific entities, Diakopoulos and Shamma (2010a) used 1820 manually labeled tweets to measure the temporal variation of sentiment, during the broadcast of U.S. presidential debate in 2008. in Asur and Huberman (2010), the authors exploited the sentiment information, in order to predict the revenue of movies after their release. Jiang et al. (2011) build a model based on 2400 manually labeled tweets and lexical features to measure the sentiment towards 5 popular queries (Obama, Google, iPad, Lakers and Lady Gaga). This approach has the benefit of accommodating different uses of words, including slang for different entities. Wang et al. (2012) was the first to build a real-time sentiment monitor,

[3] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.
[4] http://www.ranks.nl/stopwords.
[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

which was tested on the 2012 US elections. It used a model trained on 1820 manually labeled tweets and it was also the first study to also correct for humorous and sarcastic content. Finally, Mitchell et al. (2013) performed the first study that explores differences of expressed sentiment in various geographic regions.

An interesting trend that appeared in approximately 2014, was the use of Deep Neural Networks for sentiment classification (Severyn & Moschitti, 2015; Tang, Wei, Qin et al., 2014; Tang, Wei, Yang et al., 2014). In one of the first studies (Tang, Wei, Yang et al., 2014) the authors used 10 million tweets, in order to build word embeddings and achieved an impressive 86% accuracy on the task of positive vs. negative tweet classification. Kim (2014) examined a variety of Convolutional Neural Networks (CNN) and showed that most architectures even with one layer of convolution perform remarkably well.

3.3. Analysis of languages other than English

Language is a major factor in sentiment analysis, due to major differences in lexicons, syntax and semantics among languages. Today, sentiment analysis has been performed in all major languages like Spanish (Anta et al., 2013), Arabic (Duwairi et al., 2014) and Chinese (Zhao et al., 2012), all of them following language specific techniques. What is perhaps more interesting are the multilingual techniques. In cases where lexicons and text processing software is available for multiple languages, these can be combined to build multilingual systems (Tromp & Pechenizkiy, 2011). Even more challenging are language agnostic techniques. In this area, a common approach is to use emoticons as inter-language sentiment indicators (Cui et al., 2011; Narr et al., 2012). Another approach is to employ machine translation, for example Google translate (Balahur & Turchi, 2013), to 'normalize' tweets in a common language and then apply common sentiment analysis methods. When it comes to multilingual analysis in Twitter, it should be taken into consideration that the demographics of Twitter users might differ significantly between countries.

3.4. Psychometric methods

Psychometrics (Shrout & Lane, 2012) are the family of methods that attempt to assess various psychological traits of users, based on the activity and content of their online profiles. In cases where this research is focused on 'happiness', the term 'Hedonometrics' (Dodds et al., 2011) is also used. The first study in this area (Quercia et al., 2011) found statistical significant correlations between simple Twitter statistics (like number of followers) and 'The Big Five personality traits' (Barrick & Mount, 1991). For example, the number of followers was strongly associated with the 'extraversion' trait, the 'imaginative' attribute was present only to popular users and the 'organized' trait was present mainly to the influential users. The estimation of the 'big five personality traits' through textual analysis was also associated with the choice of profile picture, through image analysis (Liu et al., 2016). Another study found that happiness is assortative (Bliss et al., 2012), meaning that proximal users (distance no more than 3 links) show correlation in happiness metrics.

Another line of work is to measure emotional variation. in Pfitzner et al. (2012), the authors applied sentiment analysis techniques in a corpus of 35 million English tweets, concluding that tweets with high emotional divergence get retweeted more often. Although the polarity of the tweets does not influence the probability of retweeting, the emotional divergence does have a measurable impact. Bollen et al. (2011) measured the effect of important public events (like public holidays or general elections), on the collective sentiment. A study of Dzogang et al. (2018), analyzed 800 million tweets from the UK and measured the diurnal variation of 73 psychometric variables. The authors located two leading factors, named 'Categorical Thinking' and 'Existential Thinking', which peak at opposite time points during the 24-hour day. This study provided additional biological insights associating language use with the circadian rhythm. Emotional variation is measurable not only in the textual content of the posts, but also in the changes of the profile summaries and in the display names of Twitter users (Wesslen et al., 2018). These changes were associated with the cultural self-identity of the users. Similarly, it is possible to measure the use of certain types of language between different cultures, an area of research that belongs to *linguistic relativity*. A study that analyzed 40 million tweets, measured the emotional variation of tweets between Canada and US, confirming the stereotype that Canadians are on average more polite than Americans (Snefjella et al., 2018).

3.4.1. Health monitoring

The users' timeline in Twitter reveals practical information, not only for the psychological, but also for their physiological state. NLP techniques on Twitter have been employed for tasks like monitoring of influenza epidemic (Broniatowski et al., 2013), drug intake (Mahata et al., 2018) and incidents of intestinal disease (Zou et al., 2016), obesity and diabetes (Karami et al., 2018).

3.5. Sentiment analysis and political discourse in Twitter

Many studies have performed elaborate analyses, in order to investigate the behavior of online users, during pre-election periods. The purpose of most of these studies is to generate patterns that distinguish users' or posts' favoritism towards one political party or certain ideology. Here, the main predicament is to generate election predictions, that are close or even outperform public opinion polls (Lampos et al., 2013), to measure approval ratings (O'Connor et al., 2010), or to assess public opinion during political debates (Diakopoulos & Shamma, 2010b).

Gayo-Avello (2012) lists the major difficulties of this area that need to be addressed, before making Twitter a reliable election prediction mechanism. In brief, these difficulties are noise and demographics.

Regarding noise, a huge proportion of election-related Twitter posts are humorous, ironic or sarcastic and do not portray any party (or ideology) inclination. It is estimated that approximately half of collected tweets belong to this category (André et al., 2012; Charalampakis et al., 2015). Filtering out these posts or users is a challenging task and relies heavily on qualitative human-crafted datasets of sentiment vocabularies and pre-classified, 'ground truth' samples (Hopkins & King, 2010). Low-quality, human-curated datasets can result in a very inefficient classification algorithm, as it happened in a sarcasm detection system (González-Ibáñez et al., 2011). Existing studies on sarcasm detection are focusing on user and word selection techniques (Lampos et al., 2013), or are explicitly addressing reliability level of posts, by classifying them as rumors or trolls (Lukasik et al., 2015).

Regarding demographics, Twitter users belong to a specific social group, that is not necessarily representative of the whole electorate. Specifically, studies have indicated that Twitter users belong to a certain age (Gayo-Avello et al., 2011), social (Preotiuc-Pietro et al., 2015) and ideology demographic group and therefore, express a partial opinion of the society at best. A study of 2011 concluded that due to its demographics, Twitter is by far inferior, compared to opinion polls, for election prediction in the U.S. (Gayo-Avello et al., 2011). Another study reported that existing political party classification systems, based on sentiment analysis, are no better than random classifiers (Chung & Mustafaraj, 2011). This indicates that sentiment analysis methods are in their infancy and that they should be coupled with more sophisticated methods that incorporate rich lexical properties and context indicators, specific to each campaign (Shi et al., 2012). Fortunately, existing techniques can effectively assess and correct these biases (Preotiuc-Pietro et al., 2015).

The first term of Barack Obama's presidency (2009–2012) coincided with the immense increase of Twitter's user base and its establishment as a channel for personal political expression. As a consequence, one

Table 8
Main studies for election prediction, sarcasm detection and applications of sentiment analysis on Twitter.

Topic	Method	Reference
Election predictions	Outperform opinion polls	Lampos et al. (2013)
		Dwi Prasetyo and Hauff (2015)
	Measure approval ratings	O'Connor et al. (2010)
	Assess opinion on political debates	Diakopoulos and Shamma (2010b)
Sarcasm detection	Low-quality, human-curated dataset	González-Ibáñez et al. (2011)
	User and word selection	Lampos et al. (2013)
	Classify as rumors/trolls	Lukasik et al. (2015)
	Novel tailored lexicon	Antonakaki et al. (2017)
		Ghiassi and Lee (2018)
Sentiment analysis	Group polarization	Conover et al. (2011)
		Colleoni et al. (2014)
	Arab spring	Weber et al. (2013)
	Hugo Chavez	Morales et al. (2015)
	Climate change	Newman (2017)
		Cody et al. (2015)
	The European refugee crisis	Ross et al. (2017)
	American politics	Wang et al. (2017)

of the first studies that compared sentiment analysis in Twitter with 'traditional' opinion polls was from 2010, demonstrating a strong correlation between Sentiment Analysis in Twitter and Obama's approval ratings polls (O'Connor et al., 2010). The application of the same method in 2012 U.S. presidential elections outperformed the public opinion polls (Dwi Prasetyo & Hauff, 2015). Since then, numerous studies have performed similar analysis in other countries like Austria (Lampos et al., 2013), UK (Lampos et al., 2013) and Italy (Dwi Prasetyo & Hauff, 2015), with varying election procedures and diverse cultural and language dynamics.

Other approaches extract knowledge from the social graph, by studying the retweet or the mention graph (Conover et al., 2011), or by averaging on the predefined ideology of the political leaders that the users follow (Golbeck & Hansen, 2011; Stamatelatos et al., 2020). The tweet volume is a good indicator for a party's success, given that the correct time window is defined (Eom et al., 2015), but studies indicate that this is inefficient without sentiment analysis (Shi et al., 2012).

Sentiment analysis can measure the polarity of Twitter's users, in regard to a specific event or movement (Antonakaki, Spiliotopoulos et al., 2016; Colleoni et al., 2014; Conover et al., 2011). For example, group polarization have been studied on the context of 'Arab spring' (Weber et al., 2013), the Venezuelan president Hugo Chạvez (Morales et al., 2015), the climate change (Cody et al., 2015; Newman, 2017), the European refugee crisis (Ross et al., 2017) and American politics (Wang et al., 2017).

Other works include forming domain ontologies for sentiment analysis, like in Kontopoulos et al. (2013), where they label the tweets with sentiment score and additionally assign a sentiment grade for each notion. in Antonakaki et al. (2017) they use entities instead of ontologies, that are formed from the complete Twitter corpus for a political event in Greece, concerning the Greek bailout referendum in 2015 and the subsequent elections. This dataset contained 301,000 tweets (referendum) and 182,000 (elections). They compiled a novel political lexicon for sentiment and entity detection, tailored for the specific events. They applied volume, sentiment analysis, sarcasm correction and LDA for topic analysis.

In Table 8 we list the main studies presented here for election prediction, sarcasm detection and the application of Sentiment Analysis on Twitter.

4. Attacks and exploits

Generally, social networks have been the target of a variety of malicious attacks (Markatos et al., 2013). Here, we discuss three of the most serious categories of attacks, with high prevalence on Twitter. The first is spam, the second is automated activity from bot accounts, with the purpose of spreading misinformation and the third is hate speech.

4.1. Spam

Since its early period, Twitter has had a major problem with spam and phishing URLs. A very thorough study of 2010 (Grier et al., 2010), estimated that approximately 8% of the URLs posted in Twitter belong to one of these categories. Like in any other social network, spam on Twitter has two main properties. The first is that it is usually delivered and spread in the form of massive and orchestrated campaigns (Gao et al., 2010). The second is that spam URLs are most of the time posted from compromised (or else hijacked) accounts (Thomas et al., 2014). Therefore, two of the main research questions of this area are: Can we predict if a tweet contains spam, or if an account belongs to a spammer? What techniques do spammers employ, in order to maximize the spread of their campaigns? Perhaps the more recent review is from Wu et al. (2018), which presents the state of the art on spam detection, compares the performance of different approaches and discusses existing open issues.

4.1.1. Workflows for spam classification

Spam detection and classification are some of the most vivid research areas on Twitter. There are three types of classification tasks: (1) detecting spam tweets, (2) detecting spam users (spammers) and (3) detecting spam campaigns (Chu, Widjaja et al., 2012). If we include bot detection and sentiment classification, we realize that classification is a vital part of many Twitter studies. In Table 6, we show the general structure and main steps, commonly found in a classification workflow that uses Machine Learning methods. On the remaining of this section, we survey in detail these steps for the task of spam classification.

Labeling data as spam/legit The first part of a workflow for spam classification is the collection and labeling of tweets, according to their spam/legit status. This dataset will be used to train the classifier and assess the efficiency of the resulted classification algorithm.

Regarding collection, researchers can simply use Twitter's API to collect as many tweets as possible, expecting to 'harvest' a fair amount of spam. Interestingly, this procedure can be sped up by building 'social honeypots', where multiple legitimate accounts are set for the purpose of attracting and investigating spam and phishing URLs (Lee et al., 2010, 2011; Stringhini et al., 2010).

Regarding spam labeling, one of the most common methods is by employing human inspectors (Benevenuto et al., 2010; Flores & Kuzmanovic, 2013; Wang, 2010). Although the false positive ratio of this method is very low, an obvious disadvantage is that it requires a considerable amount of human effort. Twitter itself provides the ability to any user to report a tweet or account as spam. This method utilizes the power of the social network itself, nevertheless, reporting data has never been released by any social network.

Perhaps the most efficient method for spam labeling is through the automatic profiling of posted URLs. Due to the restricted size of messages, all URLs on Twitter are shortened to reduce their size, which also has a negative side-effect: the website that the URL points to is hidden and the user only sees the address of the shortening service along with a random identifier. Twitter has employed a URL shortener service that preemptively checks for reported malware and phishing sites before shortening a URL (Twitter Inc., 2018). A Twitter report from 2010 (Chowdhury, 2010) states that this service contributed to a drop on spam from 8% to 1%. When users click a URL posted on Twitter, they are either redirected to the initial posted URL (in case of legitimate content) or redirected to a page informing them that this URL has been flagged as malicious. This is convenient because researchers do not have to employ any sophisticated technique in order to examine the legitimacy of a link (as for example in traditional mail spam). In contrast, they only have to inspect the response of the URL shortening service when it is asked to un-shorten a URL. Example of studies that use this method for spam detection are (Amleshwaram et al., 2013) and (Gao et al., 2012).

Another method to check the validity of a URL is to query online blacklisting services (Amleshwaram et al., 2013; Antonakaki et al., 2014; Martinez-Romo & Araujo, 2013). Services that have been used on Twitter are PhishTank (Kumaraguru et al., 2007) and Google Safebrowsing (Gao et al., 2010). The major drawback of these services is that they exhibit a significant delay (it can be up to 3 days) for updating with novel malicious URLs (Sheng et al., 2009). Since Twitter is notorious for spreading information rapidly, this can be a major issue. Spammers are aware that domain blacklisting is a very efficient defense mechanism. Specifically, only 2% of spam originates from a dedicated registered domain Thomas et al. (2011). For this reason the majority of spam URLs originate from free sub-domains, such as co.cc or dot.tk. These domains cannot be blacklisted as they may contain any kind of content, including legitimate and they do not have any registration fees. Similarly spammers often exploit free blog hosting services. In the same study (Thomas et al., 2011), the authors revealed that the third most popular domain containing spam is blogspot.com. Other popular choices are LiveJournal and Wordpress. As a conclusion, domain blacklisting should be avoided, as an inefficient strategy, compared to the more targeted approach of URL blacklisting. Examples of domain blacklisting services that have been proved inefficient, due to delayed updates and containing many false positives, are URIBL and Joewein Grier et al. (2010). Similarly, traditional mail spam defense mechanisms, like Real-time Blackhole List (RBLs), are also ineffective. Another method for evading detection is multiple redirects. A spam URL that goes through multiple redirecting services and lands either in a free subdomain or even better in a blog hosting service, is the most stealthy approach.

Feature Extraction Features for spam classification are account based, like the longevity of the account, the number of posted tweets, the average tweets per day, the number of followers, the number of following and whether or not the account has a description (Chu, Widjaja et al., 2012). Some meta-features are the ratio of followers versus followings and the number of bidirectional friends. Features based on tweet content are tweet length, number of URLs posted, number of unique URLs, number of total and unique user mentions, number of trending topics and number of retweets and hasthags. URLs seem to be a valuable source of information for malevolent content. URL features include length, number of subdomains, number of redirections and age of the landing domain Aggarwal et al. (2012). Also, language features include n-grams, similarity of texts sent (a high similarity indicates that the account is actually a robot), similarity of the usernames of a user friends (Stringhini et al., 2010) and similarity between the posted trends and text (Amleshwaram et al., 2013). An interesting feature is the number of results returned from a web search of an account's name (Flores & Kuzmanovic, 2013) since fraudulent accounts rarely have a web presence. A Facebook specific feature is the user interaction graph (Gao et al., 2012), which targets users that unexpectedly interact with a high number of friends.

Usually studies extract a subset of the aforementioned features. However, there is a distinction between studies that are based on content based features (tweet text or user's profile) and graph based features (based on the properties of the social graph). Papers that belong to the first category are: Amleshwaram et al. (2013), Benevenuto et al. (2010), Lee et al. (2011), Martinez-Romo and Araujo (2013), Wang (2010) and to the second are: Amleshwaram et al. (2013), Benevenuto et al. (2010), Lee et al. (2011) and Wang (2010).

It is also possible to measure the classification ability of each feature and rank them accordingly. Available methods for this purpose are: Information gain Prusa et al. (2015), chi square (Benevenuto et al., 2010), area under the curve, Precision–Recall plot, Gini Index, Kolmogorov–Smirnov statistic, Mutual Information and Probability Ratio. Another option is to perform classification, by using a single feature and then measure this feature's accuracy.

Classification and Clustering methods After data collection, labeling and feature extraction, researchers usually feed this data to a Machine Learning algorithm and attempt to build a SPAM vs. LEGIT classifier. Various algorithms have been tested for this purpose, including Naive Bayes (Wang, 2010), Decision Trees (Gao et al., 2012; Martinez-Romo & Araujo, 2013), Random Forest (Lee et al., 2011; Mccord & Chuah, 2011), Support Vector Machines (Benevenuto et al., 2010) and Aggregate methods (Amleshwaram et al., 2013). Some studies also perform unsupervised learning (clustering), with the purpose of generating clusters based on the content (Amleshwaram et al., 2013; Gao et al., 2012; Thomas et al., 2014). Towards this direction, a very useful algorithm is the minhash (Broder, 1997).

Clustering helps grouping tweets and significantly speeds the effort of identifying spam campaigns, in collections of billions of tweets. An example of unsupervised classification for real time spam detection on Twitter is presented by Washha et al. (2019). Initially, the authors filter tweets by creating a feature vector with a predefined set of light features and then they apply periodic classification where they periodically store streamed tweets and update a labeled training dataset using unsupervised methods.

The comparison between these studies is not easy, due to the fact that they are performed in datasets collected and labeled with different methods. This brings forward the necessity for a publicly available and pre-labeled Twitter spam dataset, similar to various email spam datasets available online (Cormack, 2008). Fortunately, there are studies that have undertaken the task to compare the classification efficiency of different features and algorithms, but for the same dataset. One of these is Herzallah et al. (2018) which showed that features like the age of the account, the average time between tweets and the average length of tweets, had the best spam discriminatory ability, whereas SVM was selected as the best algorithm. Also Wu et al. (2018) compared a set of common methods from a dataset of 600 million tweets containing 6.5 million spam tweets. They concluded that most classifiers have no significant differences whereas the features containing the number of hashtags, number of user mentions and number of URLs had the highest discriminative efficiency.

Estimating the performance of a spam defense system A well-designed spam defense mechanism should have two main characteristics: accuracy and efficiency. Accuracy is measured on both sensitivity and specificity. High sensitivity means that the system correctly identifies spam content in a high ratio. High specificity means that the system has a low number of misclassified non-spam content as spam (or else false positives). In general a non-spam tweet misclassified as spam should be more penalized than a misclassified spam as a non-spam. This is because flagging or even hiding legitimate content from the user may affect more the user's overall experience from the service, than dealing with spam that eluded detection. Users are more accustomed to be exposed

Table 9
Main areas and existing studies in spam identification on Twitter.

Topic	Method	Reference
Spam classification	Campaigns	Grier et al. (2010)
	Hijacked accounts	Thomas et al. (2014)
	Blacklists drawbacks for labeling	Sheng et al. (2009) Grier et al. (2010) Martinez-Romo and Araujo (2013) Amleshwaram et al. (2013)
	Real time	Washha et al. (2019)
Clustering	Naive Bayes	Wang (2010)
	Decision Trees	Martinez-Romo and Araujo (2013) Gao et al. (2012)
	Random Forest	Lee et al. (2011)
	SVM Aggregate methods Necessity for pre-labeled datasets	Benevenuto et al. (2010)\ Amleshwaram et al. (2013) Cormack (2008)
Unsupervised learning	Cluster based on content	Amleshwaram et al. (2013) Gao et al. (2012), Thomas et al. (2014)
	Minihash	Broder (1997)
Spam defense system performance	Sensitivity/specificity: 80%, 99%	O'Donovan et al. (2012) Ozdikis et al. (2012), Grier et al. (2010)
	Time efficiency (half second)	Grier et al. (2010)
	Spammers' adaptation	Sridharan et al. (2012)
Spam techniques and practices	Bulk content distribution	Kreibich et al. (2008)
	Click-through rate	Grier et al. (2010)
	Fake accounts	Lee et al. (2011)
	Link farming	Ghosh et al. (2012)
	Trend-jacking	Martinez-Romo and Araujo (2013) Antonakaki et al. (2014) and Grier et al. (2010), Martinez-Romo and Araujo (2013)
	Account hijacking	Grier et al. (2010) Aggarwal et al. (2012) and Harvey (2010) Benevenuto et al. (2010) Almaatouq et al. (2014) Wang (2010) Lee et al. (2011) Amleshwaram et al. (2013) Flores and Kuzmanovic (2013)
Spam content	Music, games and films	Grier et al. (2010)
	Followers campaigns	Thomas et al. (2013) Stringhini et al. (2013) Perlroth (2013) Antonakaki, Polakis et al. (2016) McCoy et al. (2012) Stone-Gross et al. (2013)
	Weight loss supplements	Thomas et al. (2014)

to spam than having legit content being hidden from them. A general consensus for acceptable sensitivity and specificity values are 80% and 99%, respectively (Grier et al., 2010; O'Donovan et al., 2012; Ozdikis et al., 2012).

Efficiency measures the time overhead added to the system (or to user experience), by employing a specific defense mechanism. A very elaborate and complicated defense, regardless its success, might render a service useless, if it uses a considerable amount of time. Acceptable efficiency values are in the range of half of second. This includes both the amount of time taken to extract the features and to classify a given tweet (Grier et al., 2010).

Another consideration is that spammers are adapting quickly to avoid tracing mechanisms. Even if a defense is successful for current data, there is no evidence that the method can be robust on future spam deployed campaigns. This robustness is rarely discussed in existing studies (Sridharan et al., 2012).

4.1.2. Common spam techniques and practices

Compared to traditional email campaigns, spam on Twitter, appears to follow a more orchestrated and organized approach. Specifically, email spam relies on the bulk distribution of content towards random emails, usually harvested from web crawlers (Kreibich et al., 2008).

This is reflected in the clickthrough rate, which is the percentage of spam links that users are tricked to follow, over the sum of the total spam that they receive. For Twitter, this rate has been estimated to be 0.13% (Grier et al., 2010), which is orders of magnitude higher than the clickthrough rate of mail spam, estimated at 0.01% (Kanich et al., 2008). In this section we investigate some common exploits used by spammers for rapid content delivery.

Account hijacking Through account hijacking a single spammer can 'own' thousands of legitimate accounts and orchestrate massive spam campaigns, unbeknownst to them (Grier et al., 2010). Account hijacking on Twitter can happen either from brute force password guessing (Wisniewski, 2010) or from phishing techniques (Aggarwal et al., 2012; Harvey, 2010). It is estimated that $520 millions were lost due to phishing attempts in 2011 (Aggarwal et al., 2012). A study from 2014 (Thomas et al., 2014) revealed that out of 168 million users, 14 million had their accounts hijacked and 5 millions were deliberate fraudulent accounts. Hijacked accounts account for 69% of total spam on Twitter and the probability of users becoming victims is correlated with the number of victims that they follow. The authors also challenged one of the most profound beliefs regarding security in social media: "Only novice users can get hijacked". In contrast, they found that accounts that had many years of frequent online presence with hundreds of thousands of followers were victims as well. Social consequences are from abandoning an account (1 out of 5 victims), to losing online friends (1 out of 2 victims). This finding signifies the importance of introducing better spam defense mechanisms, as well as, raising awareness of the public on this issue and urging users to be suspicious and adopt basic practices for secure browsing.

Because of account hijacking, there is a crucial distinction between spam classification versus spammers identification (Benevenuto et al., 2010). Although spam content is very distinguishable, spam accounts can be in reality hijacked accounts that post a mix of legit and spam content. This was confirmed in a study of 100 million tweets at 2014 (Almaatouq et al., 2014), in which the authors identified two very distinct patterns of spam accounts: The first had the same tweeting and social patterns to legit users, whereas the second had in average more followings and lower betweenness centrality. Therefore, a spam message on Twitter, does not necessarily means that it was sent from a dedicated spam account. For this reason, the accuracy of tweets classification methods (such as Benevenuto et al. (2010), Wang (2010)) are usually higher than methods for account classification (such as Amleshwaram et al. (2013), Benevenuto et al. (2010), Flores and Kuzmanovic (2013), Lee et al. (2011), Wang (2010)).

In Table 9 we present the main research areas and relevant studies for fighting Spam on Twitter.

Fake accounts Another popular technique is to simply create multiple accounts and have them exhibit a tweeting pattern that attracts a fair amount of followers (Lee et al., 2011). After acquiring a critical mass of followers/targets these accounts can tweet spam content, along with harmless tweets. Of course Twitter has explicitly disallowed this practice and has built defenses against it (Twitter Help Center, 2018).

In 2018, Twitter announced (Roth & Harvey, 2018) that it has improved its spam detection techniques and as a result it suspended 70 million accounts. Twitter also 'challenges' 9.9 million accounts per week and has also posed actions against accounts that are offered as followers, in exchange for money. For example, one particular company, Devumi, has sold 200 million Twitter followers, from a collection of 3.5 million fake accounts (Confessore et al., 2018). Twitter estimated that after these actions, the average number of followers of their users will drop by 4.

Link farming and Trend-jacking One of the main objectives of spammers is to augment their targeting audience, or else to increase their followers base. To achieve this, they usually engage in activities to make a spamming account appear as 'interesting' and 'informative'. One of the most widely-used techniques is to re-post URLs to popular content such as news items, product releases and trending Internet memes. This practice is also known as 'link farming' (Ghosh et al., 2012). Other techniques to artificially increase influence are (1) adding mentions to popular users, (2) retweeting legitimate popular tweets and (3) adding hashtags with trending topics. The latter technique is called *trend-jacking*; Martinez-Romo and Araujo (2013) and Grier et al. (2010) revealed that 14% of trending topics are generated exclusively from spammers. The goal of this attack is to masquerade the spam message to make it seem innocuous and blend in with numerous other legitimate tweets about a specific topic (Antonakaki et al., 2014; Martinez-Romo & Araujo, 2013). This technique also takes advantage of the very popular and efficient search functionality of Twitter. To put this in perspective, in 2018 Google served 3.5 billion searches daily (Mangles, 2018), in 2016 Facebook's search engine received 2 billion queries per day (Constine, 2016) and in 2014, Twitter served 2 billion queries per day (Myers, 2014). Although, these companies publish usage statistics sparsely, in a way that makes it difficult to compare between each other, it is evident that searches within Twitter constitutes a significant percentage of total searches for content on the web. Therefore, 'hijacking' search results of Twitter, by mixing popular content with spam URLs, is a successful strategy.

4.1.3. Spam content

Another interesting question is *what* is the content that spammers try to promote? Access to entertainment content like music, games and films is ranked as number one (Grier et al., 2010). Interestingly content that is most often seen in email spam like pharmaceutical drugs, diet products and adult content is ranked low (less than 5% in total).

Another consideration is the rise of a fraudulent account trading marketplace, that offers additional followers. These campaigns, also called 'Gain More Follower' campaigns, attempt to attract victims by offering a mass increase to the user's number of followers. This account selling market generates $127,000–$459,000 revenue per year (Thomas et al., 2013) just by selling Twitter accounts. The same market also offers accounts for other services like Hotmail, Yahoo and Gmail. Additionally there are rough estimations of the get-more-followers type of spam, that approximate their revenue to multi-millions of dollars (Perlroth, 2013; Stringhini et al., 2013). Spammers in these campaigns follow a stealthier approach, compared to other spammers, as they manage to masquerade the malicious URLs behind legitimate and popular sites such as links to Google search results (Antonakaki, Polakis et al., 2016). The revenue of the account selling market is small compared to pharmaceutical drugs promoting campaigns, which is estimated to have a value of 185$ millions (McCoy et al., 2012) or to fake anti-virus markets with a revenue of $130 million (Stone-Gross et al., 2013). Nevertheless these markets might require to have an actual physical infrastructure (despite selling fake products), compared to the account and get-more-followers markets that require only the exploitation of account verification mechanisms of the social networks. Other popular spam content is Weight Loss Supplements and Survey Leads (Thomas et al., 2014).

4.2. Bots and the 'fake news' epidemic

Fake accounts and automatic content posting can have more dark motives than simple financial gain, as it happens with spam. Today it is considered a cultural and social phenomenon the widespread of 'news' of questionable origin and validity. This phenomenon is called the 'fake news' epidemic and is widespread on OSNs. In a recent survey, Sharma et al. (2019) reports existing methodologies, techniques and datasets regarding identification and mitigation of fake news.

Table 10
Different areas and relevant studies presented for Bots, 'fake news' and Hate speech.

Topic	Method	Reference
Bots for fake news	Russian elections	Thomas et al. (2012)
	US elections	Broniatowski et al. (2018) and Byrnes (2016)
	Australia elections	
	Measure spread	Waugh et al. (2013)
	Measure vulnerability	Vosoughi et al. (2018)
	Automatic amplification	Edwards et al. (2014)
	Honeypot	Stella et al. (2018)
	Tool: Botornot	Gilani et al. (2016) and Messias et al. (2013)
	Tool: Sybil Detector	
	Tool: Debot	Davis et al. (2016)
	Tool: RTBust	Alsaleh et al. (2014)
	Challenge: DARPA	Chavoshi et al. (2016)
	Task: Benign vs. Malign bots	Mazza et al. (2019)
	Task: Effect of elimination	Subrahmanian et al. (2016)
	Method: Anomaly detection	Chu, Gianvecchio et al. (2012)
		Shao et al. (2018)
		Rodríguez-Ruiz et al. (2020)
Rumor	Review	Meel and Vishwakarma (2019)
	Feature Analysis	O'Donovan et al. (2012)
	Feature: Retweet pattern	Mendoza et al. (2010) and O'Donovan et al. (2012)
	Feature: Reply pattern	
	Dataset: PHEME	Wu et al. (2020)
	Method: DNN	Zubiaga et al. (2016)
	Task: Astroturfing	Ajao et al. (2018)
		Ratkiewicz et al. (2011)
Hate speech	Definition and context	Fortuna and Nunes (2018)
	Definition	Waseem and Hovy (2016)
	Corpus	Founta et al. (2018)
	Yahoo! Corpus	Nobata et al. (2016)
	Feature analysis	Unsvåg and Gambäck (2018)
	Focusing on event	Burnap and Williams (2015)
	Focusing on black community	Kwok and Wang (2013)

4.2.1. Prevalence of bots and main techniques for fake content circulation

In Thomas et al. (2012), the authors studied the infrastructure used to launch a massive misinformation campaign, in order to influence political conversations regarding the outcome of 2011 Russian's parliamentary elections. The attack was done from computers around the globe, consisting of 39% of blacklisted IPs, probably originated from compromised hosts. Of course, 'opinion hijacking' through the use of Twitter bots is not only pertinent in politics. in Broniatowski et al. (2018), the authors revealed that Russian accounts that were active in US elections, were also spreading misinformation that promoted the anti-vaccination movement.

More recent studies revealed that bot activity is more wide-spread and more effective than it was thought. An analysis of 14 million tweets in 2018 demonstrated that a low number of bots (6% of total accounts) is enough to spread 31% of fake news (Shao et al., 2018). This study also revealed two of the most successful bots' strategies. The first is to reproduce low-credibility content, as early as possible, (preferably less than 10 s) after the content is originally posted. This gives the chance, for the content, to be widely spread before it is refuted. The second is to target, through user-mentions, very popular users hoping that they will retweet and redistribute the content (Stella et al., 2018). These techniques were called 'automated amplification'. Another study (Vosoughi et al., 2018) that analyzed 126,000 stories, revealed that false news-items required, in average, 10 h to reach 1500 people, whereas valid news-items required 60 h to reach the same amount of people.

Bot activities that target political campaigns are of special interest. On Twitter, hundreds of thousands of fake accounts seem to participate in orchestrated efforts to promote (or libel) certain political campaigns, an action that can be referred as 'opinion hijacking'. The phenomenon has been noticed during elections in countries like Russia (Krebs, 2018), USA (Byrnes, 2016), Australia (Waugh et al., 2013) and also in the Catalan referendum for independence (Stella et al., 2018).

4.2.2. Bot detection

Given the sophistication of bot accounts, the task of bot identification is a very challenging task. The most well-known tool that employs machine learning methods for bot detection is botornot (Davis et al., 2016). This system has been expanded and renamed to Botometer (Yang et al., 2019), using an impressive number of 1200 different features and is based on Random Forests. The Twitter Sybil Detector (Alsaleh et al., 2014) (TSD) uses Machine Learning methods on 17 features and achieves a 95% detection ratio, although it fails to detect hybrid accounts (acting both as bots and as legit), which is a main drawback, given the hijacked nature of many accounts. TSD has made publicly available a Twitter Sybils corpus that can be used for comparative analysis. DeBot (Chavoshi et al., 2016) is a detection system that exploits the fact that bots tend to post content synchronously, in contrast to humans. Similarly RTbust (Mazza et al., 2019) exploits the temporal patterns of re-tweeting bots. DARPA has challenged 6 research groups to perform bot detection for anti-vaccination campaigns (Subrahmanian et al., 2016). One of the interesting parts of the challenge was that contestants had to distinguish anti-vaccination bots from other kinds of bots. Similarly, Chu, Gianvecchio et al. (2012) tried to distinguish malevolent bots from bots that post benign content (called cyborgs, i.e. with automatic weather reports). Rodríguez-Ruiz et al. (2020) suggested a one-class classification method which attempts to simply locate deviations, or anomalies, from a 'normal' dataset . The same study also contained a thorough review and comparison analysis. The review showed that from 13 existing methods the highest efficiency is reported from Botometer for the task of identifying political tweets with the impressive AUC of 1. The comparison of 10 different Machine Learning methods, that were assigned to perform a multi-class detection from 4 different bot datasets, resulted in the Logistic Regression having the highest AUC.

Another interesting line of work is to deliberately construct a variety of harmless bots, each with different 'behavior' and then study the number of followers or other influence metrics that they acquired (Gilani et al., 2016; Messias et al., 2013). In a similar study it was

found that a group of users did not find any differences on source credibility, communication competence and interactional intentions between tweets originating from humans or bots (Edwards et al., 2014).

The defense against bots is to simply apply mechanisms for early detection and elimination. Shao et al. (2018) estimated that by eliminating only 10% of bots is enough to significantly decrease their impact. Twitter itself applies a 'quality filter' that removes possible automated content from a user's timeline (Leong, 2016).

4.2.3. Rumors and fake news detection

Bot detection is a different task than rumor detection. Although rumors and fake news exploit OSNs for rapid circulation, they do not have to be based on the existence of 'bot armies'. Most of the time, a well constructed rumor from a seemingly trustworthy source, regarding a recent and unexpected event, can be very easily propagated, even from experienced users. Consider that in 2013, a single tweet was enough for making the stock-market crash, for a short time (Matthews, 2013). One of the most thorough review containing the prevalence, consequences, datasets and classification studies is from Meel and Vishwakarma (2019).

Analysis of 18 features on a dataset between credible and non credible tweets revealed small differences (O'Donovan et al., 2012). Zubiaga et al. (2016) have studied and made freely available a dataset of 5802 tweets, regarding 5 fatal events that sparked the circulation of many fake news (also called the PHEME dataset). Analysis of this dataset with deep neural networks yielded an accuracy of 82% (Ajao et al., 2018). Another valuable source of information for rumor detection is the retweet pattern of a tweet. When combined with simple linguistic analysis, it can identify trustworthy versus invalid information spread (Mendoza et al., 2010; O'Donovan et al., 2012). Similarly, the reply pattern of a tweet has significant discriminative ability as demonstrated in a study that used DNN techniques for rumor detection (Wu et al., 2020).

An activity similar to rumor spreading, is *astroturfing*, which is the spread of positive comments, with the purpose of generating a fake 'supportive' movement towards a person or a policy (Ratkiewicz et al., 2011). Automatic defense mechanisms against the spread of fake news are just now starting to take place. Yet, we believe that the user's vigilance and well-constructed skepticism is by far the best defense.

4.3. Identification of hate speech

Hate speech in OSNs is defined as online posts and comments that are disgraceful towards individuals of certain race, religion, ethnic group or sexual orientation (Fortuna & Nunes, 2018). Perhaps the largest available corpus with hateful or abusive content in Twitter is from Founta et al. (2018), which contains 80,000 annotated tweets. Available data for hate speech detection is also available from the comments section of Yahoo! (Nobata et al., 2016) and Yahoo! Finance (Djuric et al., 2015). In Waseem and Hovy (2016), the authors defined 11 criteria for hate speech identification and developed respective NLP techniques to quantify their predictive efficiency. They also released the Twitter IDs of 16,000 manually annotated tweets. Similarly to rumor detection, common features such as profile information, network properties and text features have low discriminative ability (Unsvåg & Gambäck, 2018)

Due to the sensitive nature of this area, it is more interesting to examine the data collection methods, rather than the classification techniques which are similar to the tasks presented before. For example, one approach is to focus on specific events that can spark online debate, that might include hate speech. In a relevant study (Burnap & Williams, 2015), the authors collected 450,000 tweets containing a hashtag related to a specific event and had 2000 randomly chosen tweets, to be manually labeled as 'containing hate speech' through the CrowdFlower service. Another line of research is to identify hate speech that targets a specific group. For example, in Kwok and Wang (2013) they tested a classifier with 24,582 tweets, where half of them targeting the black community and half having neutral content.

Hate speech has been a major issue on Twitter. A study from Amnesty International (Amnesty International, 2018) reported that, on average, a woman receives an abusing tweet every 30 s and that women of color are more likely to be targets of 'troubling' tweets. Twitter has acknowledged this issue and acquired Smyte, a company that performs spam, abuse and fraud detection, with the sole purpose of addressing hate speech (Twitter Official Blog, 2018). In Table 10 we summarize the most important studies mentioned above.

5. Discussion and conclusions

Without doubt, Twitter is a fascinating OSN with some unique properties. For example: It has a typical social network structure (users connected to users), but at the same time is mostly used for news dissemination (Kantrowitz, 2018; Kwak et al., 2010; Myers et al., 2014). It supports the delivery of short text messages but at the same time it allows rich media (i.e. URLs, images, videos) and semantic (i.e. hashtags, mentions, retweets) content. It has the 1/6 of monthly active users of Facebook but the average Twitter user has four times more connections (see 2.2.2).

Although 30% of the world population is on Facebook, researchers most of the time use Twitter, in order to quickly assess the public opinion, sentiment, trend or belief regarding a subject of interest. A brief search on Google Scholar, only for 2019, will locate thousands of studies that have used Twitter, as a sampling platform for society in general. Some random examples are: Natural disaster response, social protest participation, patient safety, tobacco use, breast cancer, nonprofit communication management, corporate public relationship management, gender studies, enhancing student learning, consensus of nuclear energy policy and city planning. Given the plethora and multidisciplinary nature of this thematic potpourri, we can conclude that Twitter is an important research tool for modern science.

Yet, before we utilize and recruit Twitter as a research object, we should assess and understand its basic features, metrics, dynamics, content and dangers. This is the main objective of this survey. Initially, we presented the characteristics of the Twitter platform, along with the optimum sampling and data collection techniques. Tables 1 and 2 contains a list of the major research areas of Twitter (i.e. influence measurement and spam detection), that are based on the careful study of these characteristics (i.e. Hashtags and Retweets). In the first chapter, we also discussed the most important obstacle of the research in Twitter, which is the unavailability of gold-standard datasets, as an effect of the strict Terms of Services of the platform. In Section 2 we surveyed the social graph of Twitter. We presented studies that have attempted to extract the general structure of this graph, measured and interpreted its properties, compared it with other OSNs and assessed the temporal growth dynamics of the graph. In Section 3 we delved into studies that attempt to extract the sentiment from tweets. Sentiment is a multifaceted notion that can have simple 'positive' or 'negative' dimensions or more complex like emotion, sarcasm, humor, polarized and subjective attributes. Applying sentiment analysis on tweets with certain keywords, hashtags or geographic region can quickly give insight regarding a political campaign, a product release, or the psychological status of a group of users. Finally, in Section 4 we presented the major attacks that threaten Twitter's users. These belong to three main categories. The first are spam and phishing attempts, the second are malevolent bots that deliberately spread rumors and 'fake news' and the third is hate speech. We presented automatic detection and classification systems and we also presented the characteristics of large, massively controlled spam and 'fake news' campaigns. Given the sophistication of these attacks, we pinpoint the need to educate the public to not rely exclusively on automatic detection mechanisms, but to develop instead strong fact-checking skills.

5.1. Open questions on Twitter research

As a final remark, we should note that there are still many open or partially covered research questions. Regarding data collection and sampling, some open questions are: Given the current limitations of Twitter's API, is it possible to collect enough data to confidently measure the network properties of the social graph as it is today? The estimation of these properties requires either the complete graph of Twitter or a subset of at least 15% (Leskovec & Faloutsos, 2006) of the nodes in the graph. All the graph properties that we know today, were estimated in a time when this was practically possible. Today, obtaining a sample of that size is practically impossible even for paying options.

Moreover, given Twitter's Terms of Service, is it possible to generate and release a gold standard for various classification tasks (spam, bot, sentiment and topic identification)? Another question is, can we 'delve into the past' of Twitter and collect enough historic data, in order to study past events, as well as the temporal evolution of the social graph?

Regarding the social graph itself, some open questions that we located during this study are: Which mathematical model describes optimally the evolution of Twitter? Also, what is the relationship between Twitter's social network and the underlying real social network? Or else, can we differentiate between the accounts that we treat as news providers and the accounts that we consider as real-life 'friends'? We also noticed that some of the most basic properties of the social graph have not yet being measured confidently, or the studies that have measured them are more than five years old. Therefore some additional open questions are: Is the node degree distribution a power-law and what is the exponent of the power-law distribution? Additionally, what is the diameter of the social graph? Is the social graph steady over time, shrinking or expanding?

On sentiment analysis, we noticed that most, if not all, studies perform analysis on an ad-hoc sample of Twitter. For example, they collect all tweets that contain a set of hashtags. But is this enough to capture the public's sentiment towards an entity? If not, then what is the optimum sampling technique and size for assessing different types of sentiment? Another issue is that Twitter, as a worldwide OSN, can capture the sentiment towards various entities from different cultures and languages. Although there are many studies that perform multilingual sentiment analysis, these studies do not perform comparative analysis. Namely, they do not compare the public sentiment towards the same entity. It would be more than interesting to shed some light on the differential ways in which cultures perceive the same entity (for example the president of the USA, immigration, climate change, etc.)

Regarding spam classification and bot detection we noticed that there is some excellent research that performs very sophisticated methods for these tasks. Yet, all these studies pinpoint the need for *timely* detection. Time is of extreme essence regarding the success of spam campaigns and the spread of fake news. So, can we build a real-time bot and/or spam detector? And can we get reliable real-time indication of the trustworthiness of a tweet?

5.2. Conclusion

Twitter is constantly evolving in many ways. As a privately-owned service that seeks financial stability, as a platform that faces extreme technical challenges and as a social network that tries to fulfill the complex needs of an ever-changing user base. As an effect, the presented aspects of Twitter research should be regularly updated with surveys that cover these changes. We hope that this survey will set the ground for these future studies and help for the optimum utilization of Twitter as a research object for scientific discovery.

CRediT authorship contribution statement

Despoina Antonakaki: Conceptualization, Methodology, Investigation, Resources, Writing - original draft, Writing - review & editing, Visualization. **Paraskevi Fragopoulou:** Conceptualization, Supervision, Project administration, Funding acquisition. **Sotiris Ioannidis:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work has been supported by the European project CONCORDIA, with grant number 830927 (EUROPEAN COMMISSION Directorate-General Communications Networks, Content and Technology). Also, we would like to thank the anonymous reviewers for their meaningful and constructive comments.

References

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011a). Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended semantic web conference* (pp. 375–389). Springer.

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011b). Semantic enrichment of twitter posts for user profile construction on the social web. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, & J. Pan (Eds.), *The semanic web: Research and applications: 8th extended semantic web conference, ESWC 2011, Heraklion, Crete, Greece, May 29 – June 2, 2011, Proceedings, Part II* (pp. 375–389). Berlin, Heidelberg: Springer Berlin Heidelberg.

Abner, L. (2018). Google+ is shutting down for consumers after privacy bug. https://9to5google.com/2018/10/08/google-plus-shutting-down/. Accessed: 2018-10-27.

Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., & Gomes, J. B. (2016). A rule dynamics approach to event detection in twitter with its application to sports and politics. *Expert Systems with Applications*, *55*, 351–360.

Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: Automatic realtime phishing detection on twitter. In *2012 ECrime researchers summit* (pp. 1–12). IEEE, URL: http://ieeexplore.ieee.org/document/6489521/.

Ahmed, W. (2020). Using Twitter as a data source: an overview of social media research tools (2019). https://bit.ly/3f21WDz. Accessed: 2020-7-5.

Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society* (pp. 226–230). ACM.

Alexa Internet, Inc. (2018). Alexa top 500 global sites. http://www.alexa.com/topsites. Accessed: 2018-10-28.

Almaatouq, A., Alabdulkareem, A., Nouh, M., Shmueli, E., Alsaleh, M., Singh, V. K., Alarifi, A., Alfaris, A., & Pentland, A. S. (2014). Twitter: who gets caught? observed trends in social micro-blogging spam. In *Proceedings of the 2014 ACM conference on web science* (pp. 33–41). ACM.

Alperin, J. P., Hanson, E. W., Shores, K., & Haustein, S. (2017). Twitter bot surveys: A discrete choice experiment to increase response rates. In *#SMSociety17, Proceedings of the 8th international conference on social media & society* (pp. 27:1–27:4). New York, NY, USA: ACM.

Alsaleh, M., Alarifi, A., Al-Salman, A. M., Alfayez, M., & Almuhaysin, A. (2014). Tsd: Detecting sybil accounts in twitter. In *Machine learning and applications (ICMLA), 2014 13th international conference on* (pp. 463–469). IEEE.

Amleshwaram, A. A., Reddy, A. L. N., Yadav, S., Gu, G., & Yang, C. (2013). CATS: Characterizing automation of Twitter spammers. In *COMSNETS* (pp. 1–10). IEEE.

Amnesty International (2018). Troll patrol findings, using crowdsourcing, data science & machine learning to measure violence and abuse against women on twitter. https://bit.ly/2QAQZk9. URL: https://decoders.amnesty.org/projects/troll-patrol/findings [Online; Accessed 2018-12-30].

André, P., Bernstein, M., & Luther, K. (2012). Who gives a tweet?: Evaluating microblog content value. In *CSCW '12, CSCW '12*. New York, NY, USA: ACM.

Anta, A. F., Chiroque, L. N., Morere, P., & Santos, A. (2013). Sentiment analysis and topic detection of spanish tweets: A comparative study of of NLP techniques. *Procesamiento del Lenguaje Natural*, *50*, 45–52.

Antonakaki, D., Ioannidis, S., & Fragopoulou, P. (2018). Utilizing the average node degree to assess the temporal growth rate of Twitter. *Social Network Analysis and Mining*, *8*(1), 12.

Antonakaki, D., Polakis, I., Athanasopoulos, E., Ioannidis, S., & Fragopoulou, P. (2014). Think before rt: An experimental study of abusing twitter trends. In *International conference on social informatics* (pp. 402–413). Springer.

Antonakaki, D., Polakis, I., Athanasopoulos, E., Ioannidis, S., & Fragopoulou, P. (2016). Exploiting abused trending topics to identify spam campaigns in twitter. *Social Network Analysis and Mining*, *6*(1), 48.

Antonakaki, D., Spiliotopoulos, D., Samaras, C. V., Ioannidis, S., & Fragopoulou, P. (2016). Investigating the complete corpus of referendum and elections tweets. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 100–105). IEEE.

Antonakaki, D., Spiliotopoulos, D., V. Samaras, C., Pratikakis, P., Ioannidis, S., & Fragopoulou, P. (2017). Social media analysis during political turbulence. *PloS One*, *12*(10), Article e0186836.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (pp. 492–499). IEEE, URL: http://ieeexplore.ieee.org/document/5616710/.

Asur, S., Huberman, B. A., Szabo, G., & Wang, C. (2011). Trends in social media: Persistence and decay. *SSRN Electronic Journal*, URL: http://www.ssrn.com/abstract=1755748.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec, Vol. 10* (pp. 2200–2204).

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 3rd annual ACM web science conference on - WebSci '12* (pp. 33–42). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=2380718.2380723.

Bader, D. A., Kintali, S., Madduri, K., & Mihail, M. (2007). Approximating betweenness centrality. In *Algorithms and models for the web-graph* (pp. 124–137). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer. In *Proceedings of the fourth ACM international conference on web search and data mining - WSDM '11* (p. 65). New York, New York, USA: ACM Press.

Balahur, A., & Turchi, M. (2013). Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the international conference recent advances in natural language processing RANLP 2013* (pp. 49–55).

Barabási, A. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512, URL: http://science.sciencemag.org/content/286/5439/509.abstract.

Barbieri, N., Bonchi, F., & Manco, G. (2013). Cascade-based community detection. In *Proceedings of the sixth ACM international conference on web search and data mining* (pp. 33–42). ACM.

Barbieri, N., Bonchi, F., & Manco, G. (2014). Who to follow and why: link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1266–1275). ACM.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*(1), 1–26.

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & Society*, *30*(1), 89–116.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In *Annual collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

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

Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2013). An evolutionary algorithm approach to link prediction in dynamic social networks. CoRR abs/1304.6257.

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., & Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computer Science*, *3*(5), 388–397.

Bollen, J., Mao, H., & Pepe, A. (2010). Determining the public mood state by analysis of microblogging posts. In *Proceedings of the alife XII conference*.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.. *Icwsm*, *11*, 450–453.

Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262–278.

Bošnjak, M., Oliveira, E., Martins, J., Mendes Rodrigues, E., & Sarmento, L. (2012). Twitterecho: a distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st international conference on world wide web* (pp. 1233–1240). ACM.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System sciences (Hicss), 2010 43rd Hawaii international conference on* (pp. 1–10). IEEE.

Bray, P. (2015). Social authority: Our measure of twitter influence. http://moz.com/blog/social-authority. [Online; accessed 10-October-2015].

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117, URL: http://dl.acm.org/citation.cfm?id=297810.297827.

Broder, A. (1997). On the resemblance and containment of documents. In *SEQUENCES '97, Proceedings of the compression and complexity of sequences 1997* (p. 21). Washington, DC, USA: IEEE Computer Society, URL: http://dl.acm.org/citation.cfm?id=829502.830043.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS One*, *8*(12), Article e83672.

Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2016). Interest assortativity in twitter. In *WEBIST (1)* (pp. 239–246).

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Byrnes, N. (2016). How the Bot-y politic influenced this election. https://bit.ly/2fBN13R. Technologyreview.com, Accessed: 2018-12-30.

Carvalho, J. P., Rosa, H., Brogueira, G., & Batista, F. (2017). MISNIS: An intelligent platform for twitter topic mining. *Expert Systems with Applications*, *89*, 374–388.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010a). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining* (p. 4). ACM.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010b). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining - MDMKDD '10* (pp. 1–10). New York, New York, USA: ACM Press.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. In *4th international AAAI conference on weblogs and social media (ICWSM)*.

Cha, M., Mislove, A., & Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on world wide web - WWW '09* (p. 721). New York, New York, USA: ACM Press.

Charalampakis, B., Spathis, D., Kouslis, E., & Kermanidis, K. (2015). Detecting irony on greek political tweets: A text mining approach. In *EANN '15, Proceedings of the 16th international conference on engineering applications of neural networks (INNS)* (pp. 17:1–17:5). New York, NY, USA: ACM.

Chavoshi, N., Hamooni, H., & Mueen, A. (2016). Debot: Twitter bot detection via warped correlation. In *ICDM* (pp. 817–822).

Chen, Y., Yuan, J., You, Q., & Luo, J. (2018). Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM. arXiv preprint arXiv:1807.07961.

Choi, H.-J., & Park, C. H. (2019). Emerging topic detection in twitter stream based on high utility pattern mining. *Expert Systems with Applications*, *115*, 27–36.

Choudhury, M. D., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the 4th international AAAI conference on weblogs and social media* (pp. 34–41).

Chowdhury, A. (2010). State of Twitter Spam. https://bit.ly/2QwmB5G. Twitter.com, Accessed: 2018-12-30.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on twitter. In *International conference on applied cryptography and network security* (pp. 455–472). Springer.

Chun, H., Kwak, H., Eom, Y.-H., Ahn, Y.-Y., Moon, S., & Jeong, H. (2008). Comparison of online social relations in volume vs interaction. In *Proceedings of the 8th ACM SIGCOMM conference on internet measurement conference - IMC '08* (p. 57). New York, New York, USA: ACM Press.

Chung, J. E., & Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? In *AAAI, Vol. 11* (pp. 1770–1771).

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on twitter: an unsolicited public opinion poll. *PloS One*, *10*(8), Article e0136092.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Confessore, n., Dance, G. J., Harris, R., & Hansen, M. (2018). The follower factory. https://nyti.ms/2rJ8YZM. Accessed: 2018-10-20.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Icwsm*, *133*, 89–96.

Constine, J. (2016). Facebook sees 2 billion searches per day, but it's attacking Twitter not Google. https://tcrn.ch/2aL3jGk. Accessed: 2018-12-30.

Cormack, G. V. (2008). Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, *1*(4), 335–455, URL: http://dl.acm.org/citation.cfm?id=1454707.1454708.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695, URL: http://igraph.org.

Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Asia information retrieval symposium* (pp. 238–249). Springer.

Daniel, M., Neves, R. F., & Horta, N. (2017). Company event popularity for financial markets using Twitter and sentiment analysis. *Expert Systems with Applications*, *71*, 111–124.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web* (pp. 273–274). International World Wide Web Conferences Steering Committee.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the international conference recent advances in natural language processing RANLP 2013* (pp. 198–206).

Diakopoulos, N. A., & Shamma, D. A. (2010a). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1195–1198). ACM.

Diakopoulos, N. A., & Shamma, D. A. (2010b). Characterizing debate performance via aggregated twitter sentiment. In *CHI '10, Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1195–1198). New York, NY, USA: ACM.

Dimson, T. (2018). Emojineering part 1: Machine learning for emoji trends. https://bit.ly/2PcHKBm. Accessed: 2018-10-7.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240). ACM.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web* (pp. 29–30). ACM.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS One*, *6*(12), Article e26752.

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., & Zha, H. (2010). Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on world wide web* (pp. 331–340). ACM.

Duncan, G. (2010). It's not just you: 71 percent of tweets are ignored. https://bit.ly/2H9hZTr. [Online; accessed 2018-30-12].

Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. (2014). Sentiment analysis in arabic tweets. In *Information and communication systems (Icics), 2014 5th international conference on* (pp. 1–6). IEEE.

Dwi Prasetyo, N., & Hauff, C. (2015). Twitter-based election prediction in the developing world. In *HT '15, Proceedings of the 26th ACM conference on hypertext & social media* (pp. 149–158). New York, NY, USA: ACM.

Dzogang, F., Lightman, S., & Cristianini, N. (2018). Diurnal variations of psychometric indicators in twitter content. *PLOS ONE*, *13*(6), 1–18.

Ediger, D., Jiang, K., Riedy, J., Bader, D. A., & Corley, C. (2010). Massive social network analysis: Mining twitter for social good. In *2010 39th international conference on parallel processing* (pp. 583–593). IEEE, URL: http://ieeexplore.ieee.org/document/5599247/.

Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, *33*, 372–376.

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *SIGIR '10, Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 787–788). New York, NY, USA: ACM.

Eom, Y.-H., Puliga, M., Smailovic, J., Mozetic, I., & Caldarelli, G. (2015). Twitter-based analysis of the dynamics of collective attention to political parties. *PloS One*.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4), Article e123.

Fang, A., Macdonald, C., Ounis, I., & Habel, P. (2016). Using word embedding to evaluate the coherence of topics from Twitter data. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 1057–1060).

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with amazon's mechanical turk* (pp. 80–88). Association for Computational Linguistics.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 363–370). Association for Computational Linguistics.

Flores, M., & Kuzmanovic, A. (2013). Searching for spam: detecting fraudulent accounts via web search. In *Passive and active measurement* (pp. 208–217). Springer.

Foroozani, A., & Ebrahimi, M. (2019). Anomalous information diffusion in social networks: Twitter and Digg. *Expert Systems with Applications*, *134*, 249–266.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. arXiv preprint arXiv:1802.00393.

Freelon, D. (2018). Social media data collection tools. http://socialmediadata.wikidot.com/. Accessed: 2018-10-21.

Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, *1*.

Gabielkov, M., & Legout, A. (2012). The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM conference on CoNEXT student workshop* (pp. 19–20). ACM.

Gabielkov, M., Rao, A., & Legout, A. (2014a). Sampling online social networks: an experimental study of twitter. In *Proceedings of the 2014 ACM conference on SIGCOMM* (pp. 127–128). ACM.

Gabielkov, M., Rao, A., & Legout, A. (2014b). Studying social networks at scale: macroscopic anatomy of the twitter social graph. In *ACM SIGMETRICS performance evaluation review, Vol. 42* (pp. 277–288). ACM.

Gao, H., Chen, Y., Lee, K., Palsetia, D., & Choudhary, A. (2012). Towards online spam filtering in social networks. In *Symposium on network and distributed system security (NDSS)*.

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In *Proceedings of the 10th annual conference on internet measurement - IMC '10* (p. 35). New York, New York, USA: ACM Press.

Gayo-Avello, D. (2012). A meta-analysis of state-of-the-art electoral prediction from twitter data. CoRR abs/1206.5851. URL: http://arxiv.org/abs/1206.5851.

Gayo-Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In *ICWSM*. The AAAI Press.

Ghiassi, M., & Lee, S. (2018). A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*, *106*, 197–216.

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., & Gummadi, K. P. (2012). Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on world wide web - WWW '12* (p. 61). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=2187836.2187846.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2), 28.

Gilani, Z., Wang, L., Crowcroft, J., Almeida, M., & Farahbakhsh, R. (2016). Stweeler: A framework for twitter bot analysis. In *Proceedings of the 25th international conference companion on world wide web* (pp. 37–38). International World Wide Web Conferences Steering Committee.

Gilbert, C. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media (ICWSM-14)*.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. In *CS224N project report, Stanford, Vol. 1*.

Golbeck, J., & Hansen, D. (2011). Computing political preference among twitter followers. In *CHI '11, Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1105–1108). New York, NY, USA: ACM.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on online social networks* (pp. 27–38).

Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on twitter networks: validation of Dunbar's number. *PloS One*, *6*(8), Article e22656.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *HLT '11, Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: Short papers - volume 2* (pp. 581–586). Stroudsburg, PA, USA: Association for Computational Linguistics, URL: http://dl.acm.org/citation.cfm?id=2002736.2002850.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @spam: The underground on 140 characters or less. In *CCS '10, Proceedings of the 17th ACM conference on computer and communications security* (pp. 27–37). New York, NY, USA: ACM.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 5228–5235.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Han, B., & Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 368–378).

Han, B., Cook, P., & Baldwin, T. (2012). Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 421–432).

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. University of California Riverside.

Harvey, D. (2010). Trust and safety. https://bit.ly/2LWt3Cl. Accessed: 2018-12-30.

Hashemi, M. (2018). The infrastructure behind twitter: Scale. https://bit.ly/2qMuuJC. Accessed: 2018-10-21.

Haveliwala, T. H., & H., T. (2002). Topic-sensitive PageRank. In *Proceedings of the eleventh international conference on world wide web - WWW '02* (p. 517). New York, New York, USA: ACM Press.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. (2018). A web scraping methodology for bypassing twitter API restrictions. arXiv preprint arXiv:1803.09875.

Herzallah, W., Faris, H., & Adwan, O. (2018). Feature engineering for detecting spammers on Twitter: Modelling and analysis. *Journal of Information Science*, *44*(2), 230–247.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on world wide web - WWW '11* (p. 57). New York, New York, USA: ACM Press.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *SOMA '10, Proceedings of the first workshop on social media analytics - SOMA '10* (pp. 80–88). New York, NY, USA: ACM.

Hong, S., & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government Information Quarterly*, *29*(4), 455–461.

Hopkins, D., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Howlader, P., & Sudeep, K. (2016). Degree centrality, eigenvector centrality and the relation between them in twitter. In *2016 IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)* (pp. 678–682). IEEE.

Hu, Y., John, A., Wang, F., & Kambhampati, S. (2012). Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Twenty-sixth AAAI conference on artificial intelligence*.

Hu, X., Sun, N., Zhang, C., & Chua, T.-S. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 919–928).

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *HT '10, Proceedings of the 21st ACM conference on hypertext and hypermedia* (pp. 173–178). New York, NY, USA: ACM.

Investopedia (2020). How twitter makes money. https://tinyurl.com/y7q2cehz. [Online; accessed 2020-15-06].

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 151–160). Association for Computational Linguistics.

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2870–2879.

Johnson, S., Torres, J. J., Marro, J., & Munoz, M. A. (2010). Entropic origin of disassortativity in complex networks. *Physical Review Letters*, *104*(10), Article 108702.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., & Savage, S. (2008). Spamalytics: an empirical analysis of spam marketing conversion. In *CCS '08: Proceedings of the 15th ACM conference on computer and communications security* (pp. 3–14). New York, NY, USA: ACM, URL: http://portal.acm.org/citation.cfm?id=1455770.1455774.

Kantrowitz, A. (2018). How twitter made the tech world's most unlikely comeback. https://bit.ly/2M0sOpy. Accessed: 2018-10-21.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., & Shaw, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, *38*(1), 1–6.

Kaufmann, M., & Kalita, J. (2010). Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India, Vol. 16*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kleinberg, J. (2000). Navigation in a small world. *Nature*, *406*(6798), 845.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: measurements, models, and methods. In *Computing and combinatorics* (pp. 1–17). Springer.

Kleineberg, K.-K., & Boguñá, M. (2014). Evolution of the digital society reveals balance between viral and mass media influence. *Physical Review X*, *4*, Article 031046.

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. arXiv preprint arXiv:1507.00955.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, *40*(10), 4065–4074.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, *11*(538–541), 164.

Krebs, B. (2018). Twitter bots drown out anti-kremlin tweets. URL: https://krebsonsecurity.com/tag/maxim-goncharov/ accessed: 2018-12-30.

Kreibich, C., Kanich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., & Savage, S. (2008). On the spam campaign trail. *LEET*, *8*, 1–9.

Kucher, K., Paradis, C., & Kerren, A. (2018). The state of the art in sentiment visualization. In *Computer graphics forum, Vol. 37* (pp. 71–96). Wiley Online Library.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '06* (p. 611). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1150402.1150476.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing. In *Proceedings of the SIGCHI conference on human factors in computing systems - CHI '07* (p. 905). New York, New York, USA: ACM Press.

Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., & Kustarev, A. (2012). Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 2335–2338). ACM.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web - WWW '10* (p. 591). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1772690.1772751.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *AAAI*.

Laflin, P., Mantzaris, A. V., Ainley, F., Otley, A., Grindrod, P., & Higham, D. J. (2013). Discovering and validating influence in a dynamic online social network. *Social Network Analysis and Mining*, *3*(4), 1311–1323.

Lampos, V., Preoţiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from social media. In *ACL '13, Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 993–1003). URL http://www.aclweb.org/anthology/P13-1098.

Lee, K., Caverlee, J., & Webb, S. (2010). Uncovering social spammers. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval - SIGIR '10* (p. 435). New York, New York, USA: ACM Press.

Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*.

Leong, E. (2016). New ways to control your experience on Twitter. https://bit.ly/2b2dtRD. Accessed: 2018-12-30.

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Fourth international AAAI conference on weblogs and social media* (pp. 90–97). URL: http://arxiv.org/abs/1003.2664. arXiv:1003.2664.

Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining - KDD 08* (p. 462). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1401890.1401948.

Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '06* (p. 631). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1150402.1150479.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceeding of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining - KDD '05* (p. 177). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1081870.1081893.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. In *TKDD, Vol. 1* (p. 2). ACM.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on world wide web - WWW '08* (p. 695). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1367497.1367591.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., & Lee, B.-S. (2012). Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 721–730). ACM.

Liu, K.-L., Li, W.-J., & Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis.. In *Aaai, Vol. 12* (pp. 22–26).

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. H. (2016). Analyzing personality through social media profile picture choice.. In *ICWSM* (pp. 211–220).

LiveStats, I. (2018). Twitter usage statistics - Internet live stats. www.internetlivestats.com/twitter-statistics/. Accessed: 2018-12-30.

Lo, S. L., Chiong, R., & Cornforth, D. (2017). An unsupervised multilingual approach for online social media topic identification. *Expert Systems with Applications*, *81*, 282–298.

Lukasik, M., Cohn, T., & Bontcheva, K. (2015). Estimating collective judgement of rumours in social media. CoRR abs/1506.00468. URL: http://arxiv.org/abs/1506.00468.

Madduri, K., Ediger, D., Jiang, K., Bader, D. A., & Chavarria-Miranda, D. (2009). A faster parallel algorithm and efficient multithreaded implementations for evaluating betweenness centrality on massive datasets. In *2009 IEEE international symposium on parallel & distributed processing* (pp. 1–8). IEEE, URL: http://ieeexplore.ieee.org/document/5161100/.

Maharani, W., & Gozali, A. A. (2014). Degree centrality and eigenvector centrality in twitter. In *Telecommunication systems services and applications (TSSA), 2014 8th international conference on* (pp. 1–5). IEEE.

Mahata, D., Friedrichs, J., Shah, R. R., & Jiang, J. (2018). Did you take the pill?-detecting personal intake of medicine from twitter. arXiv preprint arXiv:1808.02082.

Mangles, C. (2018). Search engine statistics 2018. https://bit.ly/2Bwhqva. Accessed: 2018-12-30.

Markatos, E., Balzarotti, D., Almgren, M., Athanasopoulos, E., Bos, H., Cavallaro, L., Ioannidis, S., Lindorfer, M., Maggi, F., & Minchev, Z. (2013). *The red book*. SysSec Consortium.

Marketingcharts (2013). Social networking eats up 3+ hours per day for the average American user. https://bit.ly/1mmPPhB. Accessed: 2018-12-30.

Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in twitter. *Natural Language Engineering*, *20*(1), 1–28.

Martinez-Romo, J., & Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, *40*(8), 2992–3000.

Matthews, C. (2013). How does one fake tweet cause a stock market crash?. https://bit.ly/2FkPjEE. Times.com, Accessed: 2018-12-30.

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science* (pp. 183–192).

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mccord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In *International conference on autonomic and trusted computing* (pp. 175–186). Springer.

McCoy, D., Pitsillidis, A., Jordan, G., Weaver, N., Kreibich, C., Krebs, B., Voelker, G. M., Savage, S., & Levchenko, K. (2012). Pharmaleaks: understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the 21st USENIX conference on security symposium* (p. 1). USENIX Association.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012). On building a reusable twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 1113–1114). ACM.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Meeder, B., Karrer, B., Sayedi, A., Ravi, R., Borgs, C., & Chayes, J. (2011). We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 517–526). ACM.

Meel, P., & Vishwakarma, D. K. (2019). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, Article 112986.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis. In *Proceedings of the first workshop on social media analytics - SOMA '10* (pp. 71–79). New York, New York, USA: ACM Press.

Mersch, V. v. d. (2018). Twitter's 10 year struggle with developer relations. https://bit.ly/2TAG1YR. Accessed: 2018-10-28.

Messias, J., Schmidt, L., Oliveira, R., & Benevenuto, F. (2013). You followed my bot! transforming robots into influential users in twitter. *First Monday*, *18*(7).

Midha, A. (2014). Study: Exposure to brand tweets drives consumers to take action – both on and off Twitter. https://bit.ly/2CgY6UV. [Online; accessed 2018-30-12].

Milgram, S. (1967). The small world problem. In *Psychology today, Vol. 2*, New York (pp. 60–67).

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on internet measurement - IMC '07* (p. 29). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1298306.1298311.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS One*, *8*(5), Article e64417.

Morales, A., Borondo, J., Losada, J., & Benito, R. (2014). Efficiency of human activity on information spreading on twitter. In *Elsevier - Social networks, Vol. 39* (pp. 1–2011). Elsevier.

Morales, A., Borondo, J., Losada, J. C., & Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos. An Interdisciplinary Journal of Nonlinear Science*, *25*(3), Article 033114.

Motamedi, R., Jamshidi, S., Rejaie, R., & Willinger, W. (2020). Examining the evolution of the Twitter elite network. *Social Network Analysis and Mining*, *10*(1), 1.

Mottl, D. (2020). GetOldTweets3. https://github.com/Mottl/GetOldTweets3. [Online; Accessed 2020-7-5].

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PloS One*, *11*(5), Article e0155036.

Myers, L. (2014). What Happens in a Twitter Minute? Infographic. https://louisem.com/6267/twitter-minute-infographic. Accessed: 2018-12-30.

Myers, S. A., Sharma, A., Gupta, P., & Lin, J. (2014). Information network or social network?: The structure of the twitter follow graph. In *Proceedings of the companion publication of the 23rd international conference on world wide web companion* (pp. 493–498). International World Wide Web Conferences Steering Committee.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, *62*(5), 902–918.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th international workshop on semantic evaluation (Semeval-2016)* (pp. 1–18).

Narr, S., Hulfenhaus, M., & Albayrak, S. (2012). Language-independent twitter sentiment analysis. In *Knowledge discovery and machine learning (KDML), LWA* (pp. 12–14).

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd international web science conference* (p. 8). ACM.

Newman, M. E. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*(20), Article 208701.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Newman, T. P. (2017). Tracking the release of ipcc ar5 on twitter: Users, comments, and sources following the release of the working group i summary for policymakers. *Public Understanding of Science*, *26*(7), 815–825.

Nishi, R., Takaguchi, T., Oka, K., Maehara, T., Toyoda, M., Kawarabayashi, K.-i., & Masuda, N. (2016). Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining*, *6*(1), 26.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153). International World Wide Web Conferences Steering Committee.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on weblogs and social media*.

O'Donovan, J., Kang, B., Meyer, G., Höllerer, T., & Adalii, S. (2012). Credibility in context: An analysis of feature distributions in twitter. In *SocialCom/PASSAT* (pp. 293–301).

Omnicore (2018). Twitter by the numbers: Stats, demographics & fun facts. https://www.omnicoreagency.com/twitter-statistics/. Accessed: 2018-10-27.

Ozdikis, O., Senkul, P., & Oguztuzun, H. (2012). Semantic expansion of hashtags for enhanced event detection in Twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Palachy, S. (2018). A list of Twitter datasets and related resources. https://bit.ly/2H5P8zu. URL: https://github.com/shaypal5/awesome-twitter-data [Online; Accessed 2018-12-30].

Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., & Glimm, B. (2010). Making sense of twitter. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, & B. Glimm (Eds.), *The semantic web – ISWC 2010: 9th international semantic web conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised selected papers, Part I* (pp. 470–485). Berlin, Heidelberg: Springer Berlin Heidelberg.

Paul, I., Khattar, A., Kumaraguru, P., Gupta, M., & Chopra, S. (2019). Elites tweet? Characterizing the twitter verified user network. In *2019 IEEE 35th international conference on data engineering workshops (ICDEW)* (pp. 278–285). IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, *12*, 2825–2830.

Pepe, A., & Bollen, J. (2008). Between conjecture and memento: Shaping a collective emotional perception of the future. In *AAAI spring symposium: Emotion, personality, and social behavior* (pp. 111–116).

Perlroth, N. (2013). Fake twitter followers become multimillion-dollar business. *The New York Times*, Accessed: 2018-12-30.

Petrović, S., Osborne, M., & Lavrenko, V. (2010). The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media* (pp. 25–26).

Pfitzner, R., Garas, A., & Schweitzer, F. (2012). Emotional divergence influences information spreading in twitter. *ICWSM*, *12*, 2–5.

Potts, C. (2011). Sentiment symposium tutorial: Lexicons. https://bit.ly/2smM9Zo. URL: http://sentiment.christopherpotts.net/lexicons.html. [Online; Accessed 2018-12-30].

Pratikakis, P. (2018). TwAwler: A lightweight twitter crawler. arXiv preprint arXiv:1804.07748.

Preotiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying user income through language, behaviour and affect in social media. *PloS One*, URL: http://research.microsoft.com/apps/pubs/default.aspx?id=258405.

Priyanta, S., Trisna, I. P., & Prayana, N. (2019). Social network analysis of twitter to identify issuer of topic using pagerank. *International Journal of Advanced Computer Science and Applications*, *10*(1), 107–111.

Prusa, J. D., Khoshgoftaar, T. M., & Dittman, D. J. (2015). Impact of feature selection techniques for tweet sentiment classification. In *The twenty-eighth international flairs conference*.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, security, risk and trust (PASSAT) and 2011 IEEE third inernational conference on social computing (SocialCom), 2011 IEEE third international conference on* (pp. 180–185). IEEE.

Räbiger, S., & Spiliopoulou, M. (2015). A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications*, *42*(5), 2824–2834.

Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. In *Conference on weblogs and social media (ICWSM 2011)*. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850.

Reiss, J. (1981). *Statistical methods for rates and proportions* (2nd ed.). (pp. 212–225). New York: John Wiley and Sons.

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, *52*(5), 949–975.

Ritter, A., Clark, S., & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1524–1534). Association for Computational Linguistics.

Rizzo, G., & Troncy, R. (2011). Nerd: A framework for evaluating named entity recognition tools in the web of data. In *10th international semantic web conference (ISWC'11), Demo Session, Bonn, Germany* (pp. 1–4).

Rodríguez-Ruiz, J., Mata-Sánchez, J. I., Monroy, R., Loyola-González, O., & López-Cuevas, A. (2020). A one-class classification approach for bot detection on twitter. *Computers & Security*, *91*, Article 101715.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. In *Proceedings of the 20th international conference companion on world wide web - WWW '11* (p. 113). New York, New York, USA: ACM Press.

Rosa, H., Batista, F., & Carvalho, J. P. (2014). Twitter topic fuzzy fingerprints. In *2014 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 776–783). IEEE.

Rosa, H., Carvalho, J. P., & Batista, F. (2014). Detecting a tweet's topic within a large number of portuguese twitter trends. In *3rd symposium on languages, applications and technologies*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Rosen, A., & Ihara, I. (2018). Giving you more characters to express yourself. https://bit.ly/2fQ2b7W. Twitter.com, Accessed: 2018-12-30.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. arXiv preprint arXiv:1701.08118.

Roth, Y., & Harvey, D. (2018). How twitter is fighting spam and malicious automation. https://bit.ly/2N40umE. Accessed: 2018-10-20.

Sadikov, E., & Martinez, M. M. M. (2009). Information propagation on Twitter. In *CS322 project report*.

Said, A., Bowman, T. D., Abbasi, R. A., Aljohani, N. R., Hassan, S.-U., & Nawaz, R. (2019). Mining network-level properties of Twitter altmetrics data. *Scientometrics*, *120*(1), 217–235.

Saif, H., He, Y., & Alani, H. (2012a). Alleviating data sparsity for twitter sentiment analysis. In *2nd workshop on making sense of microposts (#MSM2012): Big things come in small packages at the 21st international conference on theworld wide web (WWW'12)* (pp. 2–9). CEUR Workshop Proceedings (CEUR-WS.org), URL: http://oro.open.ac.uk/38501/.

Saif, H., He, Y., & Alani, H. (2012b). Semantic sentiment analysis of twitter. In *International semantic web conference* (pp. 508–524). Springer.

Seo, Y.-D., Kim, Y.-G., Lee, E., & Baik, D.-K. (2017). Personalized recommender system based on friendship strength in social network services. *Expert Systems with Applications*, *69*, 135–148.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 959–962). ACM.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 21.

Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An empirical analysis of phishing blacklists. In *Proceedings of sixth conference on email and anti-spam (CEAS)*.

Shi, L., Agarwal, N., Agrawal, A., Garg, R., & Spoelstra, J. (2012). Predicting US primary elections with Twitter. https://stanford.io/2shORiz. accessed: 2018-12-30.

Shrout, P., & Lane, S. (2012). Psychometrics. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). New York, NY: Guilford Press.[Google Scholar].

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: article downloads, twitter mentions, and citations. *PloS One*, *7*(11), Article e47523.

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: a systematic review. *American Journal of Public Health*, *107*(1), e1–e8.

Smith, A., & Anderson, M. (2018). Social media use in 2018. https://pewrsr.ch/2FDfiFd. Accessed: 2018-12-30.

Snefjella, B., Schmidtke, D., & Kuperman, V. (2018). National character stereotypes mirror language use: A study of canadian and American tweets. *PLOS ONE*, *13*(11), 1–37.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263). Association for Computational Linguistics.

Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the first workshop on unsupervised learning in NLP* (pp. 53–63). Association for Computational Linguistics.

Sridharan, V., Shankar, V., & Gupta, M. (2012). Twitter games: How successful spammers pick targets. In *ACSAC '12, Proceedings of the 28th annual computer security applications conference* (pp. 389–398). New York, NY, USA: ACM.

Stamatelatos, G., Gyftopoulos, S., Drosatos, G., & Efraimidis, P. S. (2020). Revealing the political affinity of online entities through their twitter followers. *Information Processing & Management*, *57*(2), Article 102172.

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, Article 201803470.

Stone-Gross, B., Abman, R., Kemmerer, R. A., Kruegel, C., Steigerwald, D. G., & Vigna, G. (2013). The underground economy of Fake Antivirus Software. In *Economics of information security and privacy III* (pp. 55–78). New York, NY: Springer New York.

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference* (pp. 1–9). ACM.

Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., & Zhao, B. Y. (2013). Follow the green: growth and dynamics in twitter follower markets. In *Proceedings of the 2013 conference on internet measurement conference* (pp. 163–176). ACM.

Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., & Menczer, F. (2016). The darpa twitter bot challenge. arXiv preprint arXiv:1601.05140.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE second international conference on social computing* (pp. 177–184). IEEE.

Talukdar, P. P., & Crammer, K. (2009). New regularized algorithms for transductive learning. In *Springer Berlin Heidelberg* (pp. 442–457). Springer Berlin Heidelberg.

Tang, D., Wei, F., Qin, B., Liu, T., & Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 208–212).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers), Vol. 1* (pp. 1555–1565).

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Teevan, J., Ramage, D., & Morris, M. R. (2011). # twitterSearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 35–44). ACM.

Telegraph, M. (2018). Twitter to remove 'like' tool in a bid to improve the quality of debate. https://bit.ly/2yExMmK. Accessed: 2018-11-15.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS One*, *8*(5), Article e64841.

Thomas, K., Grier, C., & Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In *Presented as part of the 5th USENIX workshop on large-scale exploits and emergent threats*.

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: An analysis of twitter spam. In *IMC '11, Proceedings of the 2011 ACM SIGCOMM conference on internet measurement conference* (pp. 243–258). New York, NY, USA: ACM.

Thomas, K., Li, F., Grier, C., & Paxson, V. (2014). Consequences of connectivity. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security - CCS '14* (pp. 489–500). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=2660267.2660282.

Thomas, K., McCoy, D., Grier, C., Kolcz, A., & Paxson, V. (2013). Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *Proceedings of the 22nd usenix security symposium*.

Titcomb, J. (2018). Twitter makes first profit in 12-year history. https://bit.ly/2RD1MtD. telegraph.co.uk, Accessed: 2018-11-15.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. In *JSTOR - sociometry* (pp. 425–443). JSTOR.

Tromble, R., Storz, A., & Stockmann, D. (2017). We don't know what we don't know: When and how the use of twitter's public APIs biases scientific inference. Available at SSRN 3079927.

Tromp, E., & Pechenizkiy, M. (2011). Senticorr: Multilingual sentiment analysis of personal correspondence. In *Data mining workshops (ICDMW), 2011 IEEE 11th international conference on* (pp. 1247–1250). IEEE.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*(4), 402–418.

Twitter (2020). Twitter API access that scales with you and your solution. https://developer.twitter.com/en/pricing. [Online; accessed 2020-15-06].

Twitter Help Center (2018). The twitter rules. https://bit.ly/2j9xU9n. Accessed: 2018-12-30.

Twitter Inc. (2018). Shutting down spammers. https://bit.ly/2VEEZx1. Twitter.com, Accessed: 2018-12-30.

Twitter official API documentation (2018). Standard API rate limits per window. https://bit.ly/2REDPCl. Accessed: 2018-11-15.

Twitter Official Blog (2018). Continuing our commitment to health. https://bit.ly/2tocAOi. URL: https://bit.ly/2tocAOi [Online; Accessed 2018-12-30].

Twitter official blog (2018). Delivering a consistent twitter experience. https://bit.ly/2C8KX00. Accessed: 2018-11-15.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503.

Unsvåg, E. F., & Gambäck, B. (2018). The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd workshop on abusive language online (ALW2)* (pp. 75–85).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151, arXiv:http://science.sciencemag.org/content/359/6380/1146.full.pdf. URL: http://science.sciencemag.org/content/359/6380/1146.

Wang, A. H. (2010). Don't follow me - spam detection in twitter. In *SECRYPT* (pp. 142–151).

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations* (pp. 115–120). Association for Computational Linguistics.

Wang, T., Chen, Y., Zhang, Z., Sun, P., Deng, B., & Li, X. (2011). Unbiased sampling in directed social graph. *ACM SIGCOMM Computer Communication Review*, *41*(4), 401–402.

Wang, Y., Feng, Y., Hong, Z., Berger, R., & Luo, J. (2017). How polarized have we become? a multimodal classification of trump followers and clinton followers. In *International conference on social informatics* (pp. 440–456). Springer.

Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J., & Feng, X. (2014). Hashtag graph based topic model for tweet mining. In *Data mining (ICDM), 2014 IEEE international conference on* (pp. 1025–1030). IEEE.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93).

Washha, M., Qaroush, A., Mezghani, M., & Sèdes, F. (2019). Unsupervised collective-based framework for dynamic retraining of supervised real-time spam tweets detection model. *Expert Systems with Applications*, *135*, 129–152.

Waugh, B., Abdipanah, M., Hashemi, O., Abdul Rahman, S., & Cook, D. M. (2013). The influence and deception of twitter: the authenticity of the narrative and slacktivism in the Australian electoral process. In *ECCWS2014-Proceedings of the 13th European conference on cyber warefare and security*. Security Research Institute, Edith Cowan University.

Weber, I., Garimella, V. R. K., & Batayneh, A. (2013). Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 290–297). ACM.

Weitzel, L., Quaresma, P., & de Oliveira, J. P. M. (2012). Measuring node importance on twitter microblogging. In *Proceedings of the 2nd international conference on web intelligence, mining and semantics* (pp. 1–7).

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank. In *Proceedings of the third ACM international conference on web search and data mining - WSDM '10* (p. 261). New York, New York, USA: ACM Press, URL: http://dl.acm.org/citation.cfm?id=1718487.1718520.

Wernicke, S., & Rasche, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics (Oxford, England)*, *22*(9), 1152–1153, URL: http://bioinformatics.oxfordjournals.org/content/22/9/1152.long.

Wesslen, R., Nandu, S., Eltayeby, O., Gallicano, T., Levens, S., Jiang, M., & Shaikh, S. (2018). *Bumper stickers on the twitter highway: Analyzing the speed and substance of profile changes*. SocArXiv, URL: osf.io/preprints/socarxiv/bx9rm.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of facebook research in the social sciences. *Perspectives on Psychological Science*, *7*(3), 203–220.

Wisniewski, C. (2010). Twitter hack demonstrates the power of weak passwords. https://bit.ly/2sgQsFi. Accessed: 2018-12-30.

Wong, J. C., & Solon, O. (2018). Google to shut down google+ after failing to disclose user data leak. *The Guardian*, URL: https://www.theguardian.com/technology/2018/oct/08/google-plussecurity-breach-wall-street-journal.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on world wide web - WWW '11* (p. 705). New York, New York, USA: ACM Press.

Wu, Z., Pi, D., Chen, J., Xie, M., & Cao, J. (2020). Rumor detection based on propagation graph neural network with attention mechanism. *Expert Systems with Applications*, Article 113595.

Wu, T., Wen, S., Xiang, Y., & Zhou, W. (2018). Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security*, *76*, 265–284.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 177–186). ACM.

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61.

Ye, S., & Wu, S. (2010). Measuring message propagation and social influence on Twitter.com. In *International conference on social informatics* (pp. 216–231). Springer.

Yu, J., & Muñoz-Justicia, J. (2020). Free and low-cost twitter research software tools for social science. *Social Science Computer Review*, Article 0894439320904318.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining lexicon-based and learning-based methods for twitter sentiment analysis, Vol. 89*: *Technical Report HPL-2011*, HP Laboratories.

Zhao, J., Dong, L., Wu, J., & Xu, K. (2012). Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1528–1531). ACM.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338–349). Springer.

Zou, B., Lampos, V., Gorton, R., & Cox, I. J. (2016). On infectious intestinal disease surveillance using social media content. In *Proceedings of the 6th international conference on digital health conference* (pp. 157–161). ACM.

Zubiaga, A., Liakata, M., & Procter, R. (2016). Learning reporting dynamics during breaking news for rumour detection in social media. arXiv preprint arXiv:1610.07363.