XLIII CILAMCE
21-25, November, 2022
FOZ DO IGUAÇU – PR – BRAZIL

A Survey Of Machine Learning Based Techniques For Hate Speech Detection On Twitter.

Felipe R. Oliveira[1], Victoria D. Reis[1], Nelson F. F. Ebecken[1]

[1]*Dept. of Civil Engineering, Federal University of Rio de Janeiro*
Pedro Calmon, 550, 20210-030, Rio de Janeiro, Brazil
felipe.oliveira@coc.ufrj.br.

Abstract. The use of the internet and social networks, in particular for communication, has significantly increased in recent years. Twitter is the third most popular worldwide Online Social Network (OSN) only after Facebook and Instagram. Compared to others OSN's, Twitter presents a simpler data model and more straightforward data access API, which makes it a useful tool to study and analyze online behavior, including abusive patterns. This survey is an attempt to create a machine learning based guide for hate speech automatic classification including a description of twitter's technology and terminology, social graphs, sentiment analysis concepts and hate speech identification. This study also adopted a systematic literature review on the most advanced computing techniques involved with the subject, focusing on the machine learning state-of-art and research directions.

Keywords: Twitter, hate speech, machine learning, classification, sentiment analysis.

1 Introduction

Twitter is the third most popular worldwide Online Social Network (OSN) with the largest number of users around the world, around 396.5 million, behind only Facebook (2.9 billion) and Instagram (1 billion). Twitter has around 330 million monthly active users (of which 19.05 are Brazilian, the fourth country in user numbers), and an average of 500 million tweets per day [1]. The high engagement rate makes Twitter an extremely effective means of disseminating information and advertising, as well as promoting real-time social interaction between simultaneous users.

Twitter has a simple data grant model that allows the implementation of an easily scalable information acquisition infrastructure. Besides having a typical structure of OSN (with users connected to users), Twitter is mostly used for dissemination of news [2]. This particularity happens because users representing public and private institutions, news agencies, public figures, musical bands, political parties and other collectives of various natures use this social network actively as a means of dissemination and interaction with the general public. These entities, along with accounts representing individual users, make Twitter an unique subject of research in various areas such as data science, sociology, and psychology.

1.1 Terminologies

A Twitter post is a short message, called a tweet, which presents a maximum of 280 characters. A tweet can contain images, external links to other sites and videos. All accounts are, by default, public, which means that any user can read the account's tweets. A user can follow any other public user. A user's timeline contains the time series of tweets from the users he follows. Therefore, all users have a set of followers (users who receive a specific user's tweets) and followers (users whose tweets appear on a specific user's timeline).

A user can set their profile privacy settings to public or protected. Protected user's tweets are only visible to users who are pre-approved by the original sender. In addition, users on Twitter can create lists with

accounts of other users or subscribe to a list created by another user. This configuration results in a more focused (skewed) view of the timeline because it will only contain tweets originating from users belonging to this list.

In addition to texts, images, videos and external links, a tweet can contain hashtags and user mentions. In association, these items can provide metrics that allow the prediction of sociocultural trends. Lastly, users can like (or favorite) a tweet, although this feature does not necessarily reflect the user's engagement with the social network.

1.1.1. Hashtag and trending topics

A hashtag is a word preceded by a '#' character, ('#HelloWorld', for example). These words are indexed separately and users can query the platform to find tweets with specific hashtags. Hashtags have evolved into a social phenomenon and its use has been adopted by various social media (online and offline) as a simple and practical method for signifying and conceptualizing an expression in a short message [3].

The hashtags provide the following evaluation metrics: (i) frequency, which measures the number of users and posts that contain a specific hashtag, (ii) specificity, which measures the semantic relationship between the hashtag as a word and the context for which is used, (iii) consistency, which measures the level of propagation of the hashtag in different communities, and (iv) stability, which measures how the hashtag maintains its frequency and thematic content.

Popular hashtags and common search terms are listed separately as trend topics. In the literature, trend topics are also referred to as 'topics', 'popular trends' or 'popular topics'. Trends vary by geographic region and popular user interest topics. The study of Twitter trends provides valuable insights into the (i) importance, (ii) duration, and (iii) impact of real-world events. For example, an interesting question is whether Twitter is a new content generator or simply reproduces content from external sources. Studies demonstrate that Twitter performs as a content aggregator, driving specific trends to popularity [4]. Furthermore, there is a qualitative difference between trends that arise from user activities and traditional headlines that are posted by mainstream media. Specifically, events that appear first as Twitter trend topics are usually reported by individual users (accidents, demonstrations, happenings, etc.), in contrast to political events that are mainly covered by professional reporters.

1.1.2. Retweets, mentions, answers and external links

Users can retweet (repost) a tweet from another user. A user can also explicitly refer to another user, adding a mention in a tweet, using the character '@' followed by the name of the user to be quoted ('@felipe', for example). referred (retweeted or mentioned) user is notified by the service. The number of retweets is commonly associated with the value of a specific tweet's content, while the number of mentions is associated with user popularity [5].

In cases where a tweet contains an external link promoting an upcoming event, the proportion of retweets before and after the event is a good predictor of its success. For example, Asur and Huberman [6] used these metrics to predict the success of films shortly after their release. Eysenbach [7] made accurate predictions of the number of citations that a scientific article would have, based on the number of tweets containing links to online versions of the same article.

The total number of mentions users receive is associated with the influence of their profile as a whole, not the impact of their individual tweets. This is because mentions require active involvement (engagement), in contrast to simple visualization. For this reason, mentions of a user are commonly used to measure their degree of influence. The number of mentions is also useful in measuring the success of a paid advertising campaign on Twitter.

Geer's research [8] estimates that 23% of tweets already posted got at least one response. The replies generate a thread, like those typically seen in online forums. The response network is usually represented by a graph, where the nodes represent the users and the edges correspond to the response events, in a defined period of time. Another representation of the response network is the tree response cascade, which simply represents the discussion topic initiated by a single tweet, as shown by Nishi et al [9]. The shape of this tree is highly dependent on the degree (number of followers) of the root node. The number of responses a user receives is also an

influence metric. Response networks and mention networks can be used to measure the pattern of information diffusion for certain events (or hashtags). Information diffusion measures the temporal variation of the network as the information propagates.

Wu et al. [10] have shown that external links posted by media organizations become obsolete extremely quickly, while blogger links have a longer lifespan, especially when linked to music or video content. The same study also found that 50% of links posted on Twitter are generated by a very small number of companies.

1.2 Data acquisition

Since its foundation, Twitter has provided a Programming Interface Application (API) for data acquisition. Initially, Twitter used to be very open about its data access policy, but fearing that third-party services could misuse the API and build applications that can essentially mimic its website, Twitter began in 2012 to enforce stricter rules. rigid. API requests must be authenticated via OAuth2 within 15 minute windows. API requests are limited according to their type. For example, to request a user's timeline, the analyst can make 900 requests per time window. Each request can fetch up to 3,200 of the most recent tweets at a time [11] .

To overcome these limitations, some researchers use the API from fake accounts, running the risk of violating Twitter's terms of service and suspending their accounts. In addition to using the Twitter API that has the limitations presented above, another option is to build scripts that mimic the actions of a browser when visiting the main page of Twitter. This technique is called scraping, which employs advanced methods of accessing the web and is itself an area of research. TwAwler [12] is an example of a scraper capable of acquiring tweets and other metadata from a community as large as the Greek one (about 330,000 members) using a single authenticated user and common hardware.

Researchers should be aware that disclosing Twitter data is a breach of terms of service and is subject to banning from the site. Due to this punitive practice, well-studied datasets such as the Edinburgh Twitter Corpus [13] and SNAP [14] are no longer publicly available. The unavailability of public Twitter data has a serious effect on measuring the reproducibility of current surveys. Currently, there are two alternative approaches to circumvent these limitations: (i) obtaining anonymized and highly processed data releases and (ii) making available only the unique IDs of the tweets, which allows researchers to obtain the rest of the data through their own means.

The Twitter API returns data in JSON format, with a relatively complex structure. Therefore, researchers show a preference for structuring in databases that natively support information architected in that format, such as MongoDB, instead of performing complex conversions necessary for relational databases (SQL, for example).

2 Social Graphs

The social graph of an OSN represents in its vertices (or nodes) the users and in its edges the relationships between them. That is, if the user A follows the user B, this is represented in the graph with a directed edge of the node A to the node B.

Social graphs have been the center of attention in numerous fields of research. The properties of these graphs are indicative of social networks nature, portraying how users perceive the platform and how they interact with other users. They also provide insight into the dynamics of the platform and the well-being of the service. Studies on social graphs on Twitter are separated into two main categories: (i) study at the node level, aiming to produce metrics of influence, popularity and social impact of individual users, and (ii) study of the graph as a whole, trying to understand the high-level structure and dynamics of the network. Fig. 1 illustrates an example of a social graph referring to news polarization in Brazil.

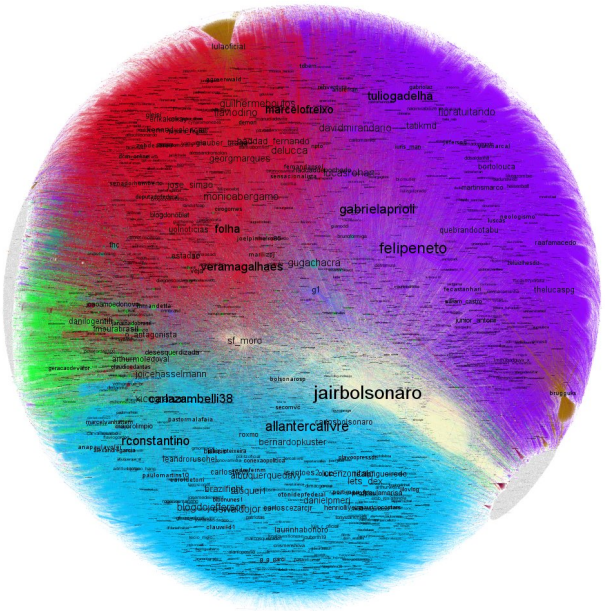

Fig. 2: Social Graph - Brazilian news polarization [56].

3 Sentiment Analysis

One of the most promising methods for content analysis on social networks is sentiment analysis. Sentiment is usually a variable that assigns values such as: positive, negative and neutral, or more specific values such as happy and sad. A single post allows for multiple sentiment assignments at once. This means that the same post can have both positive and negative feelings. In addition, additional metadata can be generated based on sentiment values. These additional features are (i) subjectivity and (ii) polarity. Subjectivity is the proportion of positive and negative tweets in relation to neutral tweets. Polarity is the ratio of positive to negative tweets. For example, sentiment analysis can measure a group's attitude towards a specific issue or assess positivity as a personality trait of a single user.

3.1 Basic Steps

The usual methodology for sentiment analysis requires pre-processing and extracting lexical features from tweets. Pre-processing steps including tokenization, expansion of abbreviations, and removal of non-semantic words and other non-lexical elements such as external links and mentions have significant effects on the performance of machine learning models.

This pre-treatment of tweets is often addressed by natural language processing (NLP), which is known as text normalization. Normalization techniques include importing information from lexicons to extend abbreviations, using spell check algorithms to correct words, automatic translators to normalize sentences, and word embeddings to measure topic similarity.

Word stem filtering, tokenization and n-grams are lexical features useful in text normalization. Fortunately, there are efficient tools that perform these tasks, through programming libraries. For example, NLTK [15] for Python and MALLET [16] for Java. Lexical features also include handling of emoticons and emojis. The study by Speriosu et al. [17] demonstrates that tweets with emoticons are four times more frequent in positive posts than in negative ones. The research by Chen et al. [18] attested to the importance of emojis for detecting positive and negative feelings.

Another popular task in NLP is named entity recognition (NER), which is defined as the process of converting a generic semantic identity into a word. For example, 'President' for 'Fernando Henrique' and 'Lugar' for 'Brasília'. The main NER tools are Stanford NER [19] and TwiNER [20], the latter has been optimized to work with Twitter data.

Manual sentiment labeling in tweets is possible through two different methods: (i) through a group of experts or (ii) with crowdsourcing techniques. Crowdsourcing is the use of online platforms that allow anyone to manually label tweets, often with a small reward. The work by Mozetič et al. [21] demonstrates how the quality of manual labeling is more important than the choice of sentiment classification method. Statistical mode is a commonly used metric to measure labeling agreement among various users.

The result of all these steps is the construction of a dataset rich in parameters, which contains ideal linguistic characteristics for identifying feelings. This dataset is usually structured as a multiplication of the vector space Tx F, where T is the number of texts and F is the number of parameters. Alternatively, the extracted resources can be modeled as a graph, by importing information from the social graph, via the label propagation method [18].

3.2 Emojis

It is difficult for users to express sentiment appropriately within limited characters in tweets. Emoji, these imagetic elements, tend to be used to express sentiment due to their expressive versatility. For this reason, studies focusing on text, including emoji, have attracted much attention within the field of natural language processing. Cui et al. [22] discovered that, in the case of a particular type of emoji, in about 20 to 40 % of tweets, the sentiments of tweets text happen to be inconsistent with those of emoji.

In most sentiment analysis research of tweets, as training data, they usually collect a roughly equal amount of positive and negative tweets by using specific filters and with manual judgment. For filtering, hashtag, emojis, named entities, and other features are typically used. For example, SemEval's sentiment analysis in Twitter task [23-28], which is a representative task group on sentiment analysis, mainly focuses on high-frequency named entities and how to collect an equal amount of positive and negative tweets. These works are based on training machine-learning sentiment analysis models utilizing collected tweets, but they build filters utilizing manually judged tweets [29]. In addition, Go et al. [30] collect tweets using specific queries and create tweets set for evaluation by selecting tweets, including sentiment among them.

Dong et al. [31] filter with testing data consisting of 25% negative, 25% positive, and 50% neutral tweets. Similarly, Kouloumpis et al. [32] create tweets set for evaluation by manually collecting and judging tweets, including sentiment for specific topics. As mentioned above, there are few studies on the approach to evaluating a classifier by using tweets following the distribution of sentiment in an actual tweets stream.

Eisner et al. [33] propose a method that uses 6,000 training data with about 1,600 emoji to train emoji word embeddings (emoji2vec) based on word embeddings trained by Google News. In Chambers's [34] work, they propose a method to identify political sentiment against nation-states through tweets. Wang et al. [35] propose classifying sentiment using the long short-term memory model (LSTM) and tweets containing emoji as training data.

Additionally, as a related study on analyzing the sentiment of tweets, considering the syntactic relationship between subjects and subjective terms, Dong et al. [31] propose a method to identify the sentiment of tweets utilizing a recurrent neural network (RNN). Xiang et al. [36] propose applying a support vector machine (SVM) to tweet sets divided into topics by a topic model. Wang et al. [37] are working on tasks to identify sentiments against more than one entity, where the sentiment against each entity is identified separately.

3.3 Embedding

Traditional sentiment analysis approaches mainly focus on feature representation through a bag-of-words, term frequency-inverse document frequency (TD-IDF), and sentiment lexicon to train a sentiment classifier. Traditional techniques of feature vector representation are commonly called non-contextual or non-semantic, as they do not explicitly consider the meaning of a word in a given context. However, with the massive success of deep learning techniques, some effective neural networks (NN) were designed to generate low-dimensional contextual representations and yield promising sentiment analysis results [31, 39-40].

Since Bengio et al. [41] pioneering work, NLP research has focused on developing new techniques for feature representation of sentences and documents based on NN approaches. Word2Vec was the first embedding

technique to hit semantic similarity between words but could not identify the meaning of a sentence based on context [42]. GloVe was created to improve Word2Vec embeddings, focusing mainly on global co-occurrence count for generating word embeddings. Using Word2Vec & GloVe, it is simpler to train with application in cases such as question answering tasks, sentiment analysis, automatic summarization, and also gained popularity in word analogy, word similarity, and named entity recognition tasks. However, the main challenge with GloVe and Word2Vec is that those tools cannot differentiate the word used in different contexts. In their work, Nakov et al introduced a deep LSTM (Long short-term memory) encoder from an attentional sequence-to-sequence model trained for machine translation (MT) to contextualize word vectors (MT-LSTM/CoVe). The results were mainly satisfactory with the main limitation with CoVe vectors being considered to be the use of zero vectors for unknown words (out-of-vocabulary words) representation.

ELMo (Embeddings from Language Models) [43] and BERT (Bidioptional Encoder Representations from Transformers) [44] embeddings are two recent popular techniques that outperform many of the NLP tasks and got massive success in neural embedding techniques that represents the context in features due to the attention-based mechanism. ELMo embedding is a character-based embedding, which allows the model to capture out of vocabulary words and deep contextualized word representation by capturing syntax and semantic features of words and outperform problems such as sentiment analysis [45] and named entity recognition [46]. In advancement to contextual embedding, BERT embedding is a breakthrough in neural embedding technique built upon transformers, including the self-attention mechanism. It can represent features with the relationship between all words in a sentence. BERT outperforms state-of-the-art feature representation for a task like question answering with SQuAD [47], language modeling/sentiment classification.

In recent years, besides the use of neural word embeddings providing better vector representations of semantic information, there has been relatively little work on direct evaluations of these models. There has been previous work to evaluate various word embedding techniques [48] applied to specific tasks like word similarity or analogy and Named entity recognition [49], with the evaluation based on the obtained performance metric.

3.4 Psychometric methods

Psychometrics is the family of methods that aim to assess the psychological traits of OSN users, based on their activities and the content of their online profiles [50]. When research specifically focuses on feelings of happiness, the term Hedonometry can also be applied. Quercia et al. [51] found significant correlations between the number of Twitter followers and the personality of its users. For example, the number of followers is strongly associated with the social extraversion trait.

Another line of research is the measurement of emotional variation. The work by Pfitzner et al. [52] applied sentiment analysis techniques in a corpus of 35 million tweets in English, concluding that tweets with high emotional divergence are retweeted more frequently. While the polarity of tweets does not influence the likelihood of their propagation, emotional divergence does have a measurable impact. Chen et al. [53] measured the effect of important public events (such as holidays or general elections) on the collective sentiment.

The study by Dzogang et al. [54] analyzed 800 million UK tweets and measured the diurnal variation of 73 psychometric variables. The authors located two main factors, termed 'categorical thinking' and 'existential thinking', which peak at opposite times during the day. This study also provided additional biological insight by linking language use to the circadian cycle.

The emotional variation is measurable not only in the textual content of the posts, but also in the changes to profile summaries and Twitter usernames. These variations are associated with the cultural self-identity of users. Likewise, it is possible to measure the use of certain types of language in different cultures, an area of research that belongs to linguistic relativity. The study by Snefjella et al. [55], which analyzed 40 million tweets, measured the emotional variation of tweets between Canada and the US, confirming the stereotype that Canadians are, on average, more polite than Americans.

Another interesting application, this time in Brazilian Portuguese, is the one made by Souza et al. in which the authors analyze OSN's data in order to infer and characterize the opinion (polarity) of the Brazilians about the impeachment process in Brazil. The work used a supervised learning approach and compared three classifiers: Max Entropy, Support Vector Machine (SVM), and Multinomial Naive Bayes. In conclusion, the SVM presented the best performance for detecting the comments' polarity about the impeachment process [56].

4 Attacks and Harmful Behavior

Social networks are targets for a variety of malicious attacks and inappropriate social practices. The three most frequent categories of attacks and harmful behavior on Twitter are: (i) the spread of spam, (ii) the activity of automatic posting scripts (bots) with the aim of spreading disinformation and (iii) hate speech.

4.1 Bots and the Misinformation Epidemic

The creation of fake accounts and the automatic posting of content can be motivated by simple financial gain, as with spam. But these old practices in RSV have gained a new purpose in the cultural and social phenomenon of the diffusion of 'news' of questionable origin and validity. This phenomenon is called the widespread misinformation epidemic [57].

The analysis made by Shao et al. [58], who studied 14 million tweets in 2018, showed that a small number of bots (6% of total accounts in a location) is enough to spread 31% of fake news. This study also revealed two of the most successful bot strategies. The first is to reproduce low-credibility content as early as possible (preferably less than 10 seconds) after the subject is originally cited. Giving the content a chance to be widely disseminated before being refuted. The second is to reach, through mentions, very popular users hoping that they will unintentionally redistribute the content. Both techniques are called automatic amplification.

Given the current sophistication of fake accounts, the task of identifying bots has become very challenging. Currently, the best bot identification tools use machine learning (ML). One of the first systems for detecting fake Twitter accounts to use ML is Botornot [59]. This system was expanded and renamed Botometer [60], using the Random Forest classification model.

Identifying bots is a different task than detecting rumors. While rumors and fake news use OSNs for quick circulation, they don't necessarily require an army of bots to spread. Most of the time, a well-crafted rumor from an apparently reliable source about a recent and unexpected event can be easily propagated, even by experienced users. It is worth mentioning that in 2013, a single tweet was enough to cause the US stock market to collapse for a short period of time [61].

4.2 Hate speech identification

Hate speech in OSN is defined as online posts and comments that offend an ethnicity, race, religion, social group or sexual orientation. Currently, the largest corpus available with hateful or abusive content on Twitter is by Founta et al. [62], which contains 80,000 labeled tweets. Waseem and Hovy [63] defined 11 criteria for identifying hate speech and established specific NLP techniques to quantify its predictive efficiency, in addition to providing 16,000 manually classified tweets.

Due to the sensitive nature of this area, it is more interesting to examine data collection methods rather than classification techniques. For example, one approach is to focus on specific events that could spark an online debate and eventually generate hate speech. Burnap and Williams [64] collected 45,000 tweets containing a hashtag related to a specific event, of which 2,000 were randomly chosen to be manually classified as containing hate speech or not, through a crowdsourcing service. Another line of research is to identify hate speech that targets a specific group. For example, Kwok and Wang [65] trained a classifier with approximately 24,000 tweets, with half of the corpus directed to the black community and half with neutral content, and obtained expressive results in the identification of racist discourses.

The spread of hate speech has been a big problem on Twitter. An Amnesty International study reported that, on average, a woman receives an abusive tweet every 30 seconds, and that women of color are more likely to be targets of offensive tweets. This study shows that as a group, women of color, (black, Asian, Latin and mixed-race women) were 34% more likely to be mentioned in abusive tweets than white women. Black women were particularly affected, with 84% more chance to be mentioned in abusive tweets when compared to white women. This data included different kinds of abuse, such as sexual, physical threats, misogyny, and racial slurs [66].

Twitter recognized this problem and acquired Smyte, a company that works in the detection of spam, abuse and fraud, with the sole purpose of eliminating hate speech from the platform.

5 Conclusion

The spread of hate speech and other nocive behaviors on social networks has increased significantly in recent years. As new social media emerge and modernize, there is a need for equally sophisticated measures to combat hate speech. Given the volume and frequency of data produced on social networks, machine learning techniques to monitor potential harmful behavior become crucial.

As with other social networks, spreading hate speech has been a big problem on Twitter. Twitter recognized this problem and acquired a company working on spam, abuse, and fraud detection, to eliminate hate speech from the platform. However, even with all of Twitter's efforts, a major social and technological challenge must be overcome to eliminate harmful behavior from the platform. Therefore, works like those presented in this article are essential to remedy the theoretical debt of combating hate speech.

Twitter is constantly evolving in many ways. As a privately owned service seeking financial stability, a platform facing technical challenges, and a social network trying to meet the complex needs of an ever-changing user base. As such, Twitter-focused research methodologies should be updated regularly. It is hoped that the present work will help future studies that use Twitter as a research object for scientific discoveries.

The works presented in this article present several paths that can be followed by taking Twitter as an object of scientific study. It is worth noting that while machine learning is a viable alternative to combating hate speech on Twitter, solutions based on machine learning are much more complex than just sentiment analysis through algorithms and mathematical models, as there are equally complex tasks involved in data acquisition, storage, and labeling.

References

[1] LiveStats, “Twitter usage statistics - Internet live stats”, 2020. [Online]. Available at: www.internetlivestats.%0Acom/twitter-statistics/.
[2] Marketingcharts, “Social networking eats up 3+ hours per day for the average American user”, 2013. [Online]. Available at: https://www.marketingcharts.com/digital-26049.
[3] D. Antonakaki, P. Fragopoulou, e S. Ioannidis, “A survey of Twitter research : Data model , graph structure , sentiment analysis”, Expert Syst. Appl., vol. 164, no September 2020, p. 114006, 2021.
[4] J. Huang, K. M. Thornton, e E. N. Efthimiadis, “Conversational Tagging in Twitter”, Hypertext and Hypermedia, vol. 10, p. 173–177, 2010.
[5] A. S. Badashian e E. Stroulia, “Measuring User Influence in Twitter -The Million Follower Fallacy”, Proc. - 3rd Int. Work. CrowdSourcing Softw. Eng. CSI-SE 2016, p. 15–21, 2016.
[6] S. Asur e B. A. Huberman, “Predicting the future with social media”, Proc. - 2010 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2010, vol. 1, p. 492–499, 2010.
[7] G. Eysenbach, “Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact.”, J. Med. Internet Res., 2011.
[8] D. Geer, “It’s not just you: 71 percent of tweets are ignored”. [Online]. Available at: https://www.wired.com/2010/10/its-not-just-you-71-percent-of-tweets-are-ignored/.
[9] R. Nishi et al., “Reply trees in Twitter: data analysis and branching process models”, Soc. Netw. Anal. Min., vol. 6, no 1, p. 1–13, 2016.
[10] S. Wu, J. M. Hofman, W. A. Mason, e D. J. Watts, “Who says what to whom on twitter”, in Proceedings of the 20th International Conference on World Wide Web, WWW 2011, 2011.
[12] Twitter, “Twitter official API documentation”, 2020. [Online]. Available at: https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits.
[13] P. Pratikakis, “twAwler: A lightweight twitter crawler”, p. 1–8, 2018.
[14] B. Hachey e M. Osborne, “Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media”, in WSA ’10: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, 2010.
[15] J. Yang e J. Leskovec, “Patterns of temporal variation in online media”, in Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011, 2011.
[16] S. Bird, S. Bird, e E. Loper, “NLTK : The natural language toolkit”, Proc. ACL-02 Work. Eff. tools Methodol. Teach. Nat. Lang. Process. Comput. Linguist. 1, 2016.
[17] A. K. McCallum, “MALLET: A Machine Learning for Language Toolkit”, 2002.
[18] M. Speriosu, N. Sudan, S. Upadhyay, e J. Baldridge, “Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph”, Proc. Conf. Empir. Methods Nat. Lang. Process., 2011.
[19] Y. Chen, Q. You, J. Yuan, e J. Luo, “Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM”, in MM 2018 - Proceedings of the 2018 ACM Multimedia Conference, 2018.
[20] J. R. Finkel, T. Grenager, e C. Manning, “Incorporating non-local information into information extraction systems by Gibbs sampling”, in ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2005.
[21] C. Li et al., “TwiNER: Named entity recognition in targeted twitter stream”, in SIGIR’12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012.
[22] H. Cui, Y. Lin, and T. Utsuro, “Sentiment Analysis of Tweets by CNN utilizing Tweets with Emoji as Training Data”, PLoS One, 2018.
[23] N.Kalchbrenner, E. Grefenstette, and P. Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In Proc. 52nd ACL. 655–665.
[24] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov. 2016. SemEval2016 Task 4: Sentiment Analysis in Twitter. In Proc. 10th SemEval. 1–18.WISDOM’18, August 20th, London.
[25] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. 2013. SemEval-2013 Task2: Sentiment Analysis in Twitter. In Proc. 7th SemEval. 312320.
[26] S. Rosenthal, N. Farra, and P. Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proc. 11th SemEval. 502–518.
[27] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In Proc. 9th SemEval. 451–463.
[28] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Proc. 8th SemEval. 73–80.
[29] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. 2013. SemEval-2013 Task2: Sentiment Analysis in Twitter. In Proc. 7th SemEval. 312320.
[30] A. Go, R. Bhayani, and L. Huang. 2009. Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford 1, 2009 (2009), 12.
[31] L.Dong, F. Wei, C.Tan, D. Tang,M. Zhou, andK. Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In Proc. 52nd ACL. 49–54.
[32] E. Kouloumpis, T. Wilson, and J. Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!. In Proc. 5th ICWSM. 538–541.
[33] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description. In Proc. 4th SocialNLP. 48–54.

[34] N. Chambers, V. Bowen, E. Genco, X. Tian, E. Young, G. Harihara, and E. Yang. 2015. Identifying Political Sentiment between Nation States with Social Media. In Proc. 20th EMNLP. 65–75.
[35] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang. 2015. Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory.. In Proc. 53th ACL. 1343–1353.
[36] B. Xiang and L. Zhou. 2014. Improving Twitter Sentiment Analysis with Topicbased Mixture Modeling and Semi-supervised Training. In Proc. 52nd ACL. 434439.
[37] B. Wang, M. Liakata, A. Zubiaga, and R. Procter. 2017. TDParse: Multi-targetspecific Sentiment Recognition on Twitter. In Proc. 15th EACL, Vol. 1. 483–493.
[39] YoonKim.2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).
[40] Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In Twenty-Fourth International Joint Conference on Artificial Intelligence.
[41] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research 3, Feb (2003), 1137–1155.
[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 3111–3119.
[43] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018).
[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
[45] Jorge A Balazs, Edison Marrese-Taylor, and Yutaka Matsuo. 2018. IIIDYT at IEST 2018: Implicit Emotion Classification With Deep Contextualized Word Representations. arXiv preprint arXiv:1808.08672 (2018).
[46] Changki Lee, Yi-Gyu Hwang, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, and Myung-Gil Jang. 2006. Fine-grained namedentity recognition using conditional random fields for question answering. In Asia Information Retrieval Symposium. Springer, 581–587.
[47] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. arXiv preprint arXiv:1902.01718 (2019).
[48] Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. 2016. Word embedding evaluation and combination.. In LREC. 300–305.
[49] Mengnan Zhao, Aaron J Masino, and Christopher C Yang. 2018. A Framework for Developing and Evaluating Word Embeddings of Drug-named Entity. In Proceedings of the BioNLP 2018 workshop. 156–160.
[50] P. Shrout e S. Lane, “Handbook of research methods for studying daily life”, Choice Rev. Online, 2012.
[51] D. Quercia, M. Kosinski, D. Stillwell, e J. Crowcroft, “Our twitter profiles, our selves: Predicting personality with twitter”, in Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011, 2011.
[52] R. Pfitzner, A. Garas, e F. Schweitzer, “Emotional divergence influences information spreading in Twitter”, in ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, 2012.
[53] M. Y. Chen e T. H. Chen, “Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena”, Futur. Gener. Comput. Syst., 2019.
[54] F. Dzogang, S. Lightman, e N. Cristianini, “Diurnal variations of psychometric indicators in twitter content”, PLoS One, 2018.
[55] B. Snefjella, D. Schmidtke, e V. Kuperman, “National character stereotypes mirror language use: A study of Canadian and American tweets”, PLoS One, 2018.
[56] B. Souza, T. Almeida, and E. Nakamura, “For or Against?: Polarity Analysis in Tweets about Impeachment Process of Brazil President”, 22nd Brazilian Symposium, 2016.
[57] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, e Y. Liu, “Combating fake news: A survey on identification and mitigation techniques”, ACM Transactions on Intelligent Systems and Technology. 2019.
[58] C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, e F. Menczer, “The spread of low-credibility content by social bots”, Nat. Commun., 2018.
[59] E. Ferrara, O. Varol, C. Davis, F. Menczer, e A. Flammini, “BotOrNot: A System to Evaluate Social Bots Clayton”, arXiv Prepr. arXiv1407.5225, 2014.
[60] K. C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, e F. Menczer, “Arming the public with artificial intelligence to counter social bots”, Hum. Behav. Emerg. Technol., 2019.
[61] C. Matthews, “How does one fake tweet cause a stock market crash?”, 2013. [Online]. Available at: https://business.time.com/2013/04/24/how-does-one-fake-tweet-cause-a-stock-market-crash/.
[62] A. M. Founta et al., “Large scale crowdsourcing and characterization of twitter abusive behavior”, in 12th International AAAI Conference on Web and Social Media, ICWSM 2018, 2018.
[63] Z. Waseem e D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”, 2016.
[64] P. Burnap e M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making”, Policy and Internet, 2015.
[65] I. Kwok e Y. Wang, “Locate the hate: Detecting tweets against blacks”, in Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013, 2013.
[66] Amnesty International, “Troll Patrol”, Amnesty International Ltd, 2019.