Identifying Malicious Networks and Individuals on Twitter by Misinformation Classification and Network Analysis

Doruk Altan
Computer Science
Bilkent University
Ankara, Turkiye
dorukaltancs@gmail.com

Alperen CAN
Computer Science
Bilkent University
Ankara, Turkiye
alperencan312@gmail.com

ABSTRACT

The spread of false information affects how well a society is able to receive knowledge, which in turn affects the communities in many manners such as politics, medical industry, daily life and more. Besides, misinformation has gained acceleration due to technological advancements - especially with the rise of social media - which has made information easily accessible by connecting huge groups of individuals. During the Covid19 pandemic, spread of misinformation has also played a devastating role based on several confusions and ambiguities about the contagion and the effects of the virus. Similarly, since false rumors about Covid19 vaccines resulted in hesitation and avoidance from taking precautions, community health was affected negatively as well as the economies of countries. Therefore, in order to minimize the negative effects, it is essential to do social network analysis of the malicious networks covering the most influential people in the social media platforms. In this way, the analysis will be helpful for authorities to overcome the negative effects by identifying the key actors and facilitating the assessment of the scale of misinformation. However, it is extremely difficult to detect and assess the misinformation scale of social media posts manually, due to the complex structure of the ever-growing network. Moreover, although algorithm based techniques exist, their success could not reach a sufficient level yet. Therefore, this study aims to find a convenient method to catch the social media posts spreading misinformation to lessen the detrimental outcomes of the mislead. It will cover Twitter data related to Covid19 vaccines. The content of the tweets from the existing misinformation dataset will be subjected to polarity and subjectivity analyses to get the measure of the probability for being a misinformation, in virtue of machine learning and artificial intelligence based algorithms and libraries such as BERT and TextBlob. Following, the results will be compared with the existing misinformation dataset. Having obtained matching results, measurements regarding node-level and network-level metrics will be held to detect the most influential users, such as energizers and brokers. Ultimately, Several tweets including the given keyword will be extracted with a Python code segment, and the same method will be applied to several tweets in order to form and assess new unique misinformation datasets.

Bilkent University, December, 2022, Ankara, Turkiye

978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/1234567890

CCS CONCEPTS

• Information systems → Information retrieval → Retrieval tasks and goals → Sentiment analysis

• Computing methodologies → Artificial intelligence → Natural language processing → Lexical semantics

• Computing methodologies → Artificial intelligence → Natural language processing → Information extraction

KEYWORDS

Social Network Analysis, SNA, Misinformation Spread, Malicious Networks, Covid19, Vaccine, Sentiment Analysis, Polarity Analysis.

ACM Reference format:

Doruk Altan and Alperen CAN. 2022. Identifying Malicious Networks and Individuals on Twitter Using Misinformation by Misinformation Classification and Network Analysis . In *Proceedings of Bilkent University (İhsan Doğramacı Bilkent University). ACM, New York, NY, USA, 4 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Social media platforms such as Twitter which make it easier to communicate, create, and share user-generated content are now widely used as a significant information source. They provide consumers the tools needed to swiftly and effectively communicate and circulate information. They can also serve as a quick means of information exchange during emergencies or as a wealth of knowledge-sharing resources [1]. On the other hand, the information generated and disseminated through social media platforms is not necessarily reliable, but presents a significant challenge to the community [2]. Misinformation, which is defined as incorrect or misleading material that is conveyed either purposefully or unintentionally, is extensively and quickly propagating on social media. Users may experience intensely negative feelings, perplexity, and worry as a result of the spread of this false information [3]. However, the outcomes are not only limited with individual-level negative effects, but also community-level, covering business, healthcare and economies. One of the most notable damages was observed during the Covid19 pandemic. The misleading information having several forms such as the ways of contagion of the virus, methods to be protected and claims about the vaccines resulted in confusion in public which eventually led to social disruption. To illustrate, several doctors claiming that the vaccines are very harmful influenced millions, gave rise to anti-vaccine demonstrations, and led to hesitation and opposition toward taking precautions, which finally caused harm in public health [4]. This situation is heavily criticized by government officials, including the minister of health of Turkiye. However, the mislead about Covid19 was not just limited to health sector workers, but numerous social media users propagating the false information also plays a great role in it. Therefore, in order to reduce the spreading of misinformation, it is important for the government to detect the social media accounts having an influencer role in this misleading process. In this paper we investigate how machine learning models and network analysis tools can help identify malicious networks or actors in Twitter that are involved in spreading misinformation regarding covid 19. We implement a classifier for detecting misinformation based on the sentiment analysis of tweets done by pre-trained NLP models. Resulting classifier will be tested against existing misinformation datasets available at Kaggle.com. In order to construct the network, we scrape tweets that were tweeted during the epidemic. The users of the tweets will be the nodes of our network and the users they mention will be the interactions they are connected by. Once the network is created, we will identify influential actors and groups by using network analysis metrics and clustering algorithms. Finally, we present our results, compare the performance of different metrics and algorithms, give our justifications for the chosen approaches and discuss the indications of our observations regarding malicious network identification.

2 Background

Recent studies show that misinformation spreading has become frequent as social media become widespread, and the possible serious outcomes must be taken into account. One of the top ten trends that the world needs to be aware of right now, according to The World Economic Forum, is the quick spread of false information online [5]. Moreover, according to a recent study on false information on social media sites, 67% of users admitted to spreading false material online, whereas 94% of participants reported having observed other individuals spread false information on social media [alp-2]. In order to minimize the negative effects of it, several studies have been carried out in order to detect the fake news in social media. However, this process maintains its difficulty as different methods such as using many fake followers also helps to increase the diffusion. Furthermore, highly active malicious user accounts play an important role to propagate fake news as a powerful source, resulting in echo chambers, increasing social polarization [6]. In order to get better results, graph convolutional networks for combining information about users and their neighbors who participate in the fake news network have been used as well as graph neural networks, which have been proven as a dominant technique based on modeling common machine learning tasks [alp-7]. In addition to AI and ML techniques, network types and metrics is also a must. For example, analyzing homogeneous networks such as friendship networks, diffusion networks and credibility networks facilitates to detect and mitigate fake news, whereas analyzing heterogeneous networks with different node sets and link types provides advantages to observe the relations from different perspectives [6]. In addition, analyzing neighbor nodes which are directly connected to at least one node of the community, examining boundary nodes, and investigating core nodes - which are only connected to members within its community - is essential to understand the role of the network structure in the detection of the fake news spreader [8]. Therefore, it is undeniable that node and graph specification methods also have a significant good use. Moreover, these techniques are crucial especially during emergency situations and unpredictable events. For instance, during Covid19 pandemic, the need for a spreader

detection model for fake news has become very evident as false information regarding various aspects was pertaining to it [8]. In order to assess the scale of misinformation of the tweets, different scores are being used related to trustworthiness and believability. These scores indicate how likely the receiver of a message is to believe its sender. Furthermore, with the help of artificial intelligence and machine learning based algorithms and libraries such as BERT and TextBlob, polarity and subjectivity analysis enables us to measure the probability of being misinformation [7]. Having achieved the desired statistics, eigenvector centrality and page rank analysis maintain its importance in the network level analysis for recognizing the most influential nodes in a network. Studies also showed that the Louvain community detection model facilitated the clustering by placing fake new sources in similar classes [9].

3 Methodology

This section is dedicated to the exploration of the problem and breaks down the necessary steps in order to identify tweets and users that are involved in the misinformation spread surrounding covid-19 vaccine.

3.1 Data Gathering

The research we conducted before deciding on our topic led us to believe that we could get access to Twitter API. However, that did not turn out to be the case. Therefore, we decided to use a scraping tool that did not require any APIs. One such tool is the snscrape from the python library. This tool allows us to scrape tweets that include certain phrases, in our case it will be “covid 19 vaccine”. Sncrape also allows us to store other relevant information such as the number of times the tweet was retweeted, the name of the user, follower number of the user, etc.. In addition to the data we scraped from Twitter, we also make use of existing misinformation datasets from Kaggle to test the performance of our approach.

3.2 Identifying Misinformation

One of the most challenging obstacles in identifying misinformation is identifying the intent of the user. In other words, whether the user intended for the tweet to be misleading or not. This makes identifying misinformation exponentially harder as we cannot solely depend on fact checking the tweets. The best models for identifying misinformation usually involve neural networks with custom layers that use many aspects of tweets such as the sentiment, polarity, truth value and content. We hope to improve and fine tune our model down the road. However, as a starting point, the main features we will be utilizing are polarity and subjectivity scores. As mentioned in section 2, many forms of misinformation such as fake news or clickbait titles often use polarizing or subjective headlines to get attention. We make the assumption that highly polarized and subjective tweets are more likely to be misinformation. The threshold values for these features will be fine tuned to give the best performance on the test set. Admittedly, polarity and subjectivity are relevant features but do not tell the whole story. To that end, we will also use credibility and/or context analysis tools in the finalized version of our model. The polarity and subjectivity scores will be generated by pre-trained models that are accessible in the Hugging Face platform. The polarity scores use the DistillBert transformer model. It had 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT’s performance. The subjectivity scores, on the other hand, are based on a model available in Hugging Face that was developed by Cloudera Fast Forward Labs which is an applied machine learning research group. The model is a version of bert-base-uncased that has been fine-tuned on a parallel corpus of 180,000 biased and neutral sentence pairs.

3.3 Network Creation

In order to detect malicious groups in a network, we have to create a network of users and form some type of connection between them. We do not have access to scrape the followers of a user, we can only see the number of followers they have. The same situation applies to retweets as well. Thus, the interactions between users will be represented by users mentioning other users. Snscrape allows us to get the mentioned user objects from a tweet through the *mentionedUsers* attribute. The network will be generated as a user ego network with mentioned users as the connections between nodes.

3.4 Identifying Malicious Groups and Individuals

While conducting network analysis on the gathered data, we will make use of network-level and node-level metrics as well as different clustering approaches. The performance of different algorithms and metrics will be compared to each other. The model we use will be based on the best performing algorithms and metrics. In order to identify the most influential individual actors we will make use of betweenness centrality and degree centrality. High number of connections will be attributed to influential actors and betweenness metrics will help us identify the nodes that hold power by being intermediaries. For classifying malicious groups in the network, we will utilize the Louvain

community detection algorithm and the Girvan-Newman algorithm. Louvain clustering assigns scores to clusters based on how densely the nodes are connected compared to how they would be in a random network. The Girvan-Newman method for the detection and analysis of community structure is based on the iterative elimination of edges with the highest number of the shortest paths that go through them. By eliminating edges the network breaks down into smaller networks, i.e. communities [11].

ACKNOWLEDGMENTS

We would like to thank to Ms. Miray Kaş and Ms. Gözde Yazıcı.

REFERENCES

[1] Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: a rumor theoretic analysis of tweets during social crises.

[2] Chen, X. and Sin, S.C.J., 2013. 'Misinformation? What of it?' Motivations and individual differences in misinformation sharing on social media. Proceedings of the Association for Information Science and Technology.

[3] Budak, C., Agrawal, D., & Abbadi, A. E. (2011). Limiting the spread of misinformation in social networks. Paper presented at the International World Wide Web Conference, Hyderabad, India.

[4] *Misperceptions about doctors' trust in covid-19 vaccines influence vaccination rate* (2022) *Misperceptions about doctor's trust in Covid-19 vaccines influence vaccination rate* | *Max-Planck-Gesellschaft*. Available at: https://www.mpg.de/18755942/0601-pat-misperceptions-about-doctor-s-trust-in-covid-19-vaccines-influence-vaccination-rate-916457-x (Accessed: December 3, 2022).

[5] World Economic Forum. (2014). Top trends of 2014: 10.

[6] Shu, Kai & Bernard, H. & Liu, Huan. (2018). Studying Fake News via Network Analysis: Detection and Mitigation.

[7] Michail, Dimitrios & Kanakaris, Nikos & Varlamis, Iraklis. (2022). Detection of fake news campaigns using graph convolutional networks. International Journal of Information Management Data Insights. 2. 100104. 10.1016/j.jjimei.2022.100104.

[8] Rath, Bhavtosh & Salecha, Aadesh & Srivastava, Jaideep. (2020). Detecting Fake News Spreaders in Social Networks using Inductive Representation Learning.

[9] Srebrenik, B. (2019) *Ego network analysis for the detection of fake news*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/ego-network-analysis-for-the-detection-of-fake-news-da6b2dfc7c7e (Accessed: December 3, 2022).

[10] Shu, K., Bernard, H. R., & Liu, H. (2019). Studying fake news via network analysis: detection and mitigation. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 43-65). Springer, Cham.

[11] Despalatovic, Ljiljana & Vojkovic, Tanja & Vukicevic, Damir. (2014). Community structure in networks: Girvan-Newman algorithm improvement. 997-1002. 10.1109/MIPRO.2014.6859714.

[12]Möbius (2020) *Covid-19 fake news dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/arashnic/covid19-fake-news (Accessed: December 3, 2022).