Detecting and Identifying Fake News on Twitter

Lenna Nashif
Tandon School of Engineering
New York University
New York, USA
Email: lan9199@nyu.edu

Abstract— **This paper delves into the profound impact of social media on relaying information, which is often stored and hosted in the cloud. The ability to differentiate between correct information and information that can be termed "misinformation" or "fake news" is integral for social media platforms. The spread of misinformation can lead to severe and possibly negative effects. To understand this further, this paper uses Big Data Analytics, often applicable in cloud computing, cross-referenced with reliable newspaper sources, to understand a tweet's validity in the context of the Covid-19 pandemic. Tweepy and TextBlob are Python libraries that are used to extract, derive sentiment analysis and subjectivity, and critically analyze the data for trends and implications in tweets. This analysis then is used to locate where the misinformation is spreading from. Through rigorous testing and verification, it becomes possible to determine and indicate in a simple and effective way which tweets are reliable and which are not. Implementing cloud storage to build this out on a larger scale opens up the exciting possibility of applying this method of locating fake news on Twitter to other trending topics, including elections, scientific discussions, and sporting events.**

Keywords-social media; misinformation; Covid-19

I. INTRODUCTION

Social media is an integral part of contemporary society. Information is purveyed from one of many platforms divulging critical news at all hours of the day. This news can be coming from all corners of the globe. Such an interconnected display of communication, having brought extreme benefits, can also have downfalls. One such example is the spread of misinformation on all social media platforms, and specifically on Twitter. This misinformation ranges from harmless to very serious, with consequences that cause ripple effects worldwide. Twitter in particular is interesting to look at because it is used by so many people and is able to capture their attention in bite-sized, attention-grabbing statements. Though other social media platforms might be able to delve into the misinformation on a deeper level, users who share misinformation typically do so out of convenience rather than ill-intentions, which makes Twitter a more ideal platform to understand the phenomenon. Limiting this misinformation or identifying and indicating which tweets are incorrect can better educate users on the truth and protect them from some of the possible harms.

Current solutions tend only to determine a tweet's validity using machine learning algorithms based on engagement and comparing a tweet's contents to news articles [1]. However, controversial tweets with varying opinions can be inconclusive in many instances, causing many solutions to fail. Taking this into consideration, using a layered approach for validating tweets may be a more reliable solution. This paper proposes using a combination of *TextBlob* (a Python library for processing textual data) and *Tweepy* (a Python library that provides easy-to-use access to the Twitter API) to develop a more robust algorithm.

The paper's organization is as follows: Section II will cover additional works related to this topic. Section III presents the motivation behind this research. Section IV will further explain the technical implementation. The conclusions and possible future work close the article.

II. RELATED WORK

There are several different approaches to address the issue of false information on Twitter. One such technique is found in the study *In Detecting Fake News with Tweets' Properties* [1]. Fake news datasets were found online and were analyzed using the machine learning news classification algorithm and ensemble classification model. To understand, dissect, and evaluate the information, they used data mining to classify features related to fake news, using Decision Tree, Random Forest and Extra Tree Classifier [1]. This approach was met with success, with accuracy ranging from 99.8% to 44.15%. There was one caveat in that it maintained the assumption that the media is always the source of complete truth.

Similarly, Verma et al. [2] looked to approach this problem through the classification method. The naïve Bayes classifier and the passive-aggressive classifier were used to construct a prediction versus the actual matrix. A tweet was determined to be either real positive, false positive, false negative, or true negative. Afterward, a mathematical formula was used that calculated accuracy, precision, and recall outputting a final score for both methods. The passive-aggressive classifier itself produced 78% accuracy [2], which

 ISBN: 978-1-61208-845-7

overall is not bad, but the final score was 50% for both scenarios, which lends more uncertainty than is desirable.

The Nikam et al. [3] work approached this differently. Each tweet that was looked into was given an individual score after comparing it to news sources [3]. Then, an overall user score was devised as a result of different conditions such as engagement and location. The two scores of the tweet and that of the user were used to create an overall score for the tweet's reputability.

III. MOTIVATION

Twitter continues to play a leading role in the worldwide dissemination of information. A great example is the 2020 US Presidential election, when Twitter began using warning labels for posts that the company believed shared false claims about the election, as well as Covid-19. Though this has not been implemented across their entire platform, there are sometimes severe consequences of showing misinformation, showing that taking proactive steps to stop this misinformation is necessary. The prevalence of Twitter use had increased throughout the Covid-19 quarantine when large numbers of people had to stay home. Twitter provided a means for socialization when other methods of interacting, primarily in-person, were limited. This increase in Twitter users brought to the forefront the necessity of having the misinformation be highlighted. Though misinformation can be harmless, sometimes the aftermath can have detrimental effects on a person's reputation, mental health, and finances [4].

Currently, Twitter has an estimated 330 million monthly active users. According to a Pew Research Center study, Twitter users tend to be younger, more likely to identify as Democrats, more highly educated, and have higher incomes than US adults overall. Most users are passive users [5], while the top 10% most prolific accounts create 80% of all content on the platform. By making the content from these active accounts more transparent, Twitter can prevent misinformation or general confusion amongst users. In this way, users can make decisions with all of the necessary information available.

IV. HYPOTHESIS AND EMPIRICAL EVIDENCE

Tweets were pulled from Twitter using Tweepy to access Twitter's API. In this case, upwards of 10K Tweets related to Covid-19 were analyzed. TextBlob's sentiment detector was used to understand the tweet's sentiment, whether positive, negative, or neutral, since misinformation is often associated with strong opinions. A tweet's overall subjectivity was also provided through TextBlob's analysis. Additionally, it is necessary to understand how a news source cited in a tweet affects reliability. By cross-referencing the cited news source with a list of already-verified resources, it is possible to reinforce if a Tweet is reliable or not. Tweets with news sources are marked in Twitter using a macro that is triggered from Python. This list was created by looking at sources that were ranked in the middle of the political spectrum so that there is little subjectivity in the reporting. For sources that are more scientific, peer-reviewed and internationally renowned sources were deemed acceptable. Included in this list is The BBC, the Associated Press, and the World Health Organizatio. This list is a continuing work, since it is possible for sources to become more or less reliable over time.

Once these two elements, tweet sentiment and news source verification, are examined, an intuitive way of seeing the tweet's validity is implemented. Bringing all of the analysis together, the tweet is marked with one of three options: a checkmark to indicate that it is likely accurate, a question mark for a tweet that requires further investigation but does not seem immediately suspicious, and a warning sign if it appears clear that there is some misinformation.

The first step in this method of understanding tweets was to connect to Twitter's API and get the dataset. Tweepy allows for tweets, retweets (including quoted retweets), favorites, replies, and followers to be extracted. Keywords related to Covid-19, such as "coronavirus", "Covid-19", and "wearamask," were used to find specific tweets. These were then run through a Python script for sentiment analysis. Additional details were surmised through Excel, and an Excel macro was created to automate the process further. The resultant information was tabulated and graphed. The list of reliable sources was cross-referenced to sources in the tweets. Cross-referencing added an element of truthfulness and reliability to them. Finally, the tweets were evaluated on whether or not both attributes were verified.

To replicate this for other topical discussions, the initital objective is to determine relevant keywords for the topic. Twitter has a list function for certain topics that can be used for this. Then, the tweets using these keywords should be pulled using Tweepy. The tweets are run through the Python code to understand sentiment and subjectivity, both key indicators of a tweets general tone and substance. If a tweet cites a news source, the news source is double-checked against a pre-defined repository of reliable sources. Lastly, the overall reliability is evaluated, based on subjectivity, sentiment, and sources. A high degree of subjectivity and strong sentiment would indicate that the tweet might not be accurate. The accuracy of the tweet is simply shown through a checkmark, a question mark, or a warning sign.

 ISBN: 978-1-61208-845-7

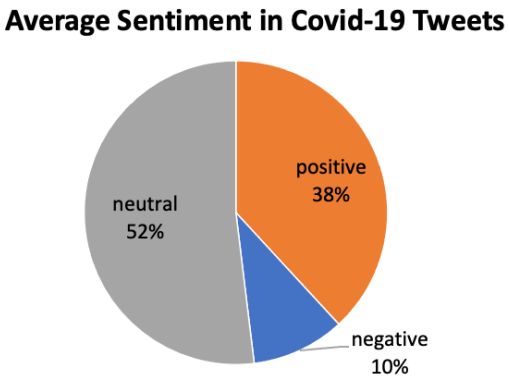

Figure 1. Results based on Textblob sentiment analysis

Figure 1 indicates the base sentiment of the tweets that were analyzed, showing that the majority of tweets, 52%, had a neutral sentiment. 38% had a positive sentiment, and only 10% of tweets regarding Covid-19 were negative.Tweets that were neutral tended to have less differentiation on whether or not they were for Covid-19 precautions. Misinformation regarding Covid-19 precautions often created divisive opinions and so being neutral showed that there was less nuance to the tweets, and therefore it was more likely to not be misinformation.

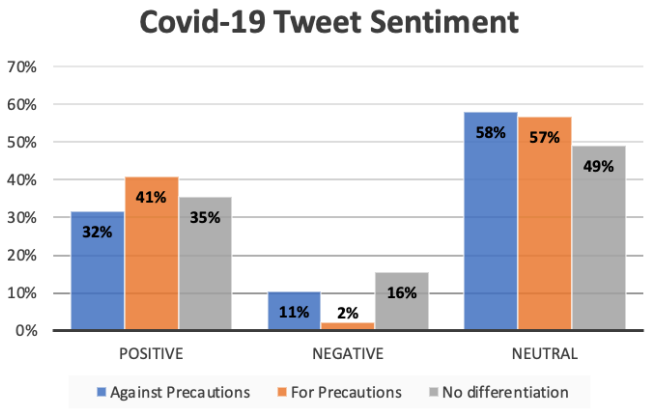

Figure 2. Additional results based on Textblob sentiment analysis.

Figure 2 delves deeper into the data shown in Figure 1. It shows that a tweet with neutral language had no clear indication of being for or against precautions. A negative tweet had a higher likelihood of being against safeguards or having no differentiation. Similarly, a tweet that was positive correlated with a greater chance of being for precautions. This can be taken a step further to say that if there is a tweet with a negative sentiment, that is against Covid-19 precautions, there is a higher chance that that tweet might have some type of misinformation

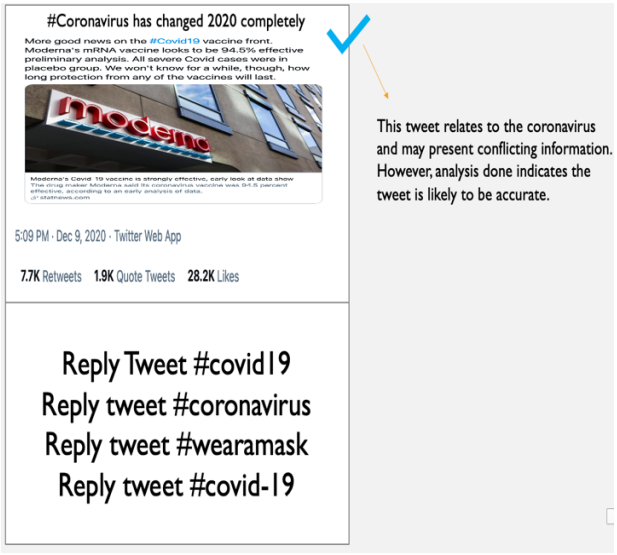

Figure 3. Sample Diagram of a Final Implementation on Twitter

Lastly, examining Figure 3 shows the finalized view on Twitter with a sample tweet that is neither for nor against precautions and citing a reliable scientific website. With the two aspects utilized, it can be confidently asserted that the tweet does not spread misinformation, as represented in the checkmark. Once a user hovers above the checkmark, there will be a blurb to explain why it is a reliable tweet.

V. Conclusion and Future Work

This paper proposed a method to bring awareness to possible Twitter misinformation based on empirical evidence, Python and Excel analysis, and the Twitter API. Tweets were evaluated to show the sentiment (positive, negative, or neutral) of a tweet that had been determined to be for or against Covid-19 precautions. Additionally, the average subjectivity for the tweets was also evaluated. Overall, the paper showed a clear method of how to identify and label a mixture of topics on Twitter as misinformation.

The efforts carried out show promising results of how to approach misinformation on Twitter. Future work will focus on cross-referencing with different news sources, additional and enhanced sentiment analysis, and analyzing larger Twitter datasets. Additionally, the current work only focused on text-based tweets, but due to the nature of Twitter, analyzing images containing text would be beneficial as well. By utilizing this method, it becomes possible to analyze tweets and data on various topics and ideas.

Acknowledgment

I would like to thank Dr. Aspen Olmsted for his support and guidance in navigating the process of writing and submitting this paper.

I would also like to thank Jason Nelson and Bansri Shah for their help in the paper-writing process.

 ISBN: 978-1-61208-845-7

REFERENCES

[1] N. X. Nyow and H. N. Chua, "Detecting Fake News with Tweets' Properties," 2019 IEEE Conference on Application, Information and Network Security (AINS), Pulau Pinang, Malaysia, 2019, pp. 24-29, doi: 10.1109/AINS47559.2019.8968706. [Accessed: 21-Nov-2020]

[2] P. K. Verma, V. Sharma, and S. Agarwal, "Credibility investigation for tweets and its users," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 925-928, doi: 10.1109/ICCMC.2019.8819809. [Accessed: 21-Nov-2020]

[3] S. S. Nikam and R. Dalvi, "Machine Learning Algorithm based model for classification of fake news on Twitter," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 1-4, doi: 10.1109/I-SMAC49090.2020.9243385. [Accessed: 21-Nov-2020]

[4] Z. Thomas, "What is the cost of 'cancel culture'?," *BBC News*, 08-Oct-2020. [Online]. Available: https://www.bbc.com/news/business-54374824. [Accessed: 23-Nov-2020].

[5] S. Wojcik and A. Hughes, "How Twitter Users Compare to the General Public," *Pew Research Center: Internet, Science & Tech*, 30-May-2020. [Online]. Available: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. [Accessed: 23-Nov-2020]

 ISBN: 978-1-61208-845-7