STEM Fellowship Big Data Challenge 2021

Investigating the role of Misinformation on Twitter through Passive-Aggressive Classification

Scott Blender[1], Kelly Ly[2], Hadassah Galapo[3], and En Yu Yap[4]

[1]Temple University, Department of Mathematics
[2]University of Massachusetts Lowell, Department of Computer Science
[3]Temple University, Department of Computer Science
[4]University of Hong Kong, Department of Chemistry

September 28, 2021

Abstract

In this paper, we consider the problem of misinformation spreading on Twitter. We collected a large dataset of more than 1.2 million tweets related to the topic of the COVID-19 vaccine from May 3rd, 2021 to May 24th, 2021. We trained a binary classification model to detect whether a tweet was considered misinformation or not misinformation. The Passive-Aggressive Classifier (PAC) outperformed logistic regression and was selected to classify the entire dataset. Overall, 79.83% of tweets were classified as not misinformation and 20.17% as misinformation. Our results also highlight the distribution of sentiment and the most popular terms in both the tweets classified as misinformation and not misinformation. Through our study, we propose a workflow to assess COVID-19-related Twitter behavior using misinformation classification, topic analysis, sentiment analysis, and retweeting behavior. From our findings, we suggest that platforms like Twitter establish policies to reduce the spread of misinformation and that public health officials inform the public of key topics that appear frequently in misinformation tweets.

Keywords
COVID-19, Twitter, Passive-Aggressive Classification, Misinformation, Machine Learning

1 Introduction

In March, President Biden launched a national strategy to allow all Americans to be eligible for the vaccine by May 1 and declared that there will be enough vaccines for every American by the end of May [1]. Currently, as the end of May approaches, 41.2% of all Americans have been fully vaccinated and present research indicates that the number is plateauing [2].

There is evidence to suggest that a declining uptake in vaccines and disparities in vaccination rates among the general public may be attributed to trust, socio-econo-demographic, and political factors, suggesting low vaccination rates and vulnerability to misinformation are related [3]. Misinformation, more commonly referred to as "fake news", is the spread of false information. Due to the extensive usage of social media platforms, a wealth of misinformation surrounding COVID-19 vaccine safety has spread around the world. Such falsehoods have influenced behaviors around mask-wearing and social distancing, and more recently, vaccination hesitancy [4]. Therefore, quantifying the amount of misinformation and understanding its source is essential in ensuring safe social media spaces.

In recent months, Twitter, a social media platform with 350 million users, has observed an unprecedented amount of misinformation surrounding the vaccine and has demanded researchers and social media platforms tackle the ongoing surge of false information [4]. This study uses Twitter data to analyze how misinformation influences public sentiment and general compliance to COVID-19 health policies and vaccinations.

2 Materials & Methods

In this study, social media data was used to classify what constitutes misinformation, in tweets, and understand the relationship between tweets

that contain misinformation and the implications misinformation on social media may have. Specifically, we assume that tweets can be decomposed into tweets that contain misinformation or do not contain misinformation [4] [5].

To obtain tweets for our study, we scraped tweets using `snscrape` (https://github.com/JustAnotherArchivist/snscrape) with the keywords "covid vaccine" in our query. We scraped 1,248,103 tweets for the dates of May 3rd, 2021 through May 24th, 2021. The information scraped with the text of the tweets includes the author's username, tweet ID, account verification status, and the detected language of the tweet. The tweets' IDs scraped are publicly available at (https://github.com/scottblender/twitter-covid-19-vaccine-analysis) and can be rehydrated for future analyses.

We used a supervised machine learning approach to classify tweets as misinformation or not misinformation. We labeled a subset of 500 tweets randomly selected from the full dataset by hand. This dataset was slightly imbalanced, with 30.2% of tweets manually assigned as misinformation, and 69.8% manually assigned as not misinformation. In the training dataset, 25.6% of tweets were assigned as misinformation, and 74.4% were assigned as not misinformation. Before using binary classification, we cleaned and tokenized the tweet text to help optimize the model's performance [6]. After dropping all the non-English tweets, the data was preprocessed: we converted all text to lowercase, removed usernames and links, removed punctuation, removed stopwords, and lemmatized each string.

Once the tweets were preprocessed, we computed the term frequency-inverse document frequency (TF-IDF) such that terms that appeared commonly in all the tweets were ignored ($\max\_df = 0.8$). For this classification problem, we used Passive Aggressive Classification (PAC) and Logistic Regression.

The Passive-Aggressive Classifier is an algorithm that updates its weights when it finds instances where its predictions are wrong. It is an online algorithm that uses one example at a time to update its weights and never sees this example again, unlike a batch algorithm. This is why PAC is useful for large datasets where examples are rapidly changing such as news articles or Tweets [7].

We ran PAC and Logistic Regression and compared model performance before applying the better performing model to a larger dataset of 1,248,103 tweets. PAC outperformed Logistic Regression with a higher F1 score of 37.7% compared to a score of 0%. In this instance, due to the F1 score of 0%, Logistic Regression acts as a baseline model, where all tweets were classified as not misinformation.

Sentiment analysis and the distribution of common words in tweets with respect to being misinformation or not misinformation were also computed. The VADER (Valence Aware Dictionary for Sentiment Reasoning) method was used for sentiment analysis.

3 Results

The PAC model achieved a model accuracy of 65.60% and an F1 score of 37.7%.

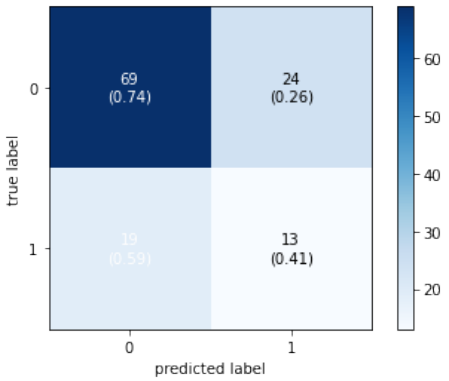

Figure 1: Confusion matrix for PAC model. Each cell displays the number of tweets the model classified in relation to manual classification results.

Figure 1 shows a confusion matrix for the PAC model. In terms of classification accuracy, 26% of tweets were wrongly classified by the PAC from the training set as being misinformation when they were not misinformation. For tweets that were manually classified as misinformation, the model wrongly classified 59% of tweets. The model was able to classify 41% of tweets that were classified as misinformation correctly and 74% of tweets that were classified as not misinformation.

Status	**Verified**	**Number of Retweets**
Misinformation	No	64478
Misinformation	Yes	53032
Not Misinformation	No	170799
Not Misinformation	Yes	149401

Table 1: Summary of results for number of retweets based on account verification and class designation.

Overall, 79.83% of tweets were classified as not misinformation and 20.17% as misinformation.

For tweets from non-verified accounts, 20.54% were classified as misinformation and 79.46% as not misinformation. For tweets from verified accounts, 15.23% were classified as misinformation and 84.77% as not misinformation.

The number of retweets from unverified accounts that do not contain misinformation is 170799. The number of retweets from unverified accounts that contain misinformation is 64478.

The number of retweets from verified accounts that do not contain misinformation is 149401. The number of retweets from verified accounts that do contain misinformation is 53032.

This means that 12.11% of retweets from verified accounts are misinformation and 14.73% of retweets from unverified accounts are misinformation.

Out of all the tweets, 41.26% were positive or overly positive, 32.7% neutral, and 26.02% were negative or overly negative.

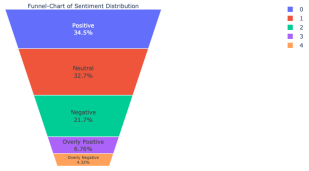

Figure 2: Funnel chart of sentiment distribution of scraped tweets.

Popular words that appeared in tweets classified as misinformation include 'victims', 'people', 'fuck', 'worse', 'asia', 'deaths', and 'certificate'. Popular words that appeared in tweets classified as not misinformation include 'covidvaccine', 'dose', 'help', 'calm', and 'shot'.

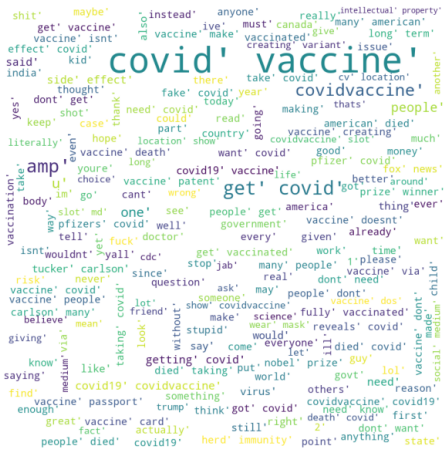

Figure 3: Word cloud for tweets classified as negative sentiment.

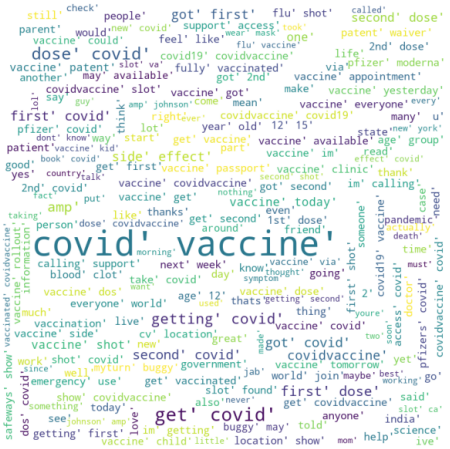

Figure 4: Word cloud for tweets classified as positive sentiment.

4 Discussion

In this study, the PAC model was used to classify tweets and identify whether or not tweets were misinformation. Key results include topics and words most common in each tweets that are misinformation or are not misinformation, which types of accounts did misinformation originate from, the number of retweets misinformation tweets received, and sentiment distributions of tweets that are either misinformation or not. While some might suggest our training dataset is imbalanced, due to time constraints, we treated this as a standard machine learning problem.

In terms of general information, 'doses' and 'shot' were common tokens or words that appeared in tweets classified as not misinformation. These tokens were commonly found in tweets that can be interpreted as status updates, with people posting about themselves receiving a dose of the vaccine. In addition, other words found in general information included '12', relating to posts about updates to age restrictions on vaccine eligibility.

'Victims' and 'people' appeared in tweets classified with misinformation related to tweets from users using hashtags such as '#covidvaccinevictims' or when talking about the side effects of the vaccine on people of different age groups. Victims appeared in tandem with tweets about misinformation circulating about how the vaccine affects menstruation for women as well. Targeted content to alleviate misinformation on association and causation may be useful in helping social media users understand the relationship between topics that appear frequently in misinformation tweets, including the aforementioned relation between COVID-19 and menstrual cycles.

Out of all verified accounts, only 11.68% sent tweets classified as misinformation by the PAC. Out of all unverified accounts, 16.67% sent tweets classified as misinformation. This lends insight into which types of accounts are more likely to spread misleading information.

A simple way for misinformation to spread is through retweets. 12.11% of retweets from verified accounts are classified as misinformation and 14.73% of retweets from unverified accounts are classified as misinformation according to the PAC. To reduce the amount of verified accounts spreading misinformation, Twitter can adopt or establish policies that limit the spread of posts from these accounts.

As seen in Figure 2, out of all the tweets scraped, 34.5% of tweets were classified as having positive sentiment. Analyzing the data further, this can be related to the keywords associated with tweets that do not contain misinformation, since a larger majority of these tweets are users providing vaccine-status updates and showing expression of contentment with new policies and lessening of restrictions. Since the majority of tweets are not misinformation, it seems plausible that the greatest fraction of tweets are positive, with the second largest fraction of tweets, 32.7%, being classified as neutral sentiment.

21.7% of tweets were classified as negative, which can also be related to tweets containing misinformation. Tweets in this category often contain words like the tokens mentioned above that are lexically more negative, such as 'fuck', 'worse', and 'victim', all of which contributed largely to the model's identification of tweets containing misinformation [8].

Additional words that contributed to sentiment classification are displayed in the word clouds in Figures 3 and 4. Figure 3 shows how some words such as 'Tucker Carlson', 'Country', or 'Government' imply that tweets with negative sentiment may contain political affirmations or associations, an outlet for extension. Figure 4 shows that many words relating to positive sentiment contain information about people 'getting covid vaccine'.

As for limitations, in this study, due to the semi-imbalanced dataset initially scraped, oversampling may have alleviated class imbalances in training data. However, since the tweets we scraped were not previously classified, limitations on how to oversample occurred, resulting in model inaccuracy in classifying the minority class. This is shown in Figure 1, where only 41% of training data is properly classified as misinformation.

Due to the limitations on time and resources, we manually classified a small, random sample of tweets. To improve future studies, additional samples can be generated and manually classified, and finally, compared against one another to determine which sample has the least class imbalance. Algorithms such as heterogeneous subpopulation identification can also be applied to determine which groups are similar for supervised algorithm such as PAC.

However, even with limited accuracy, we were able to use PAC to classify the larger dataset of tweets and extract key implications from our results.

In terms of areas of extension, seeing which topics contribute or appear frequently in misinformation tweets lends insight into how public health officials may be able to help inform the public on various topics. Further analysis can be done to see how many tweets related to topics such as the relationship between the vaccine affecting menstruation cycles come from verified accounts and how many retweets these tweets are receiving or where tweeters are located who are sharing this information.

Other potential areas for extension include using the geotagged nature of certain tweets to determine the source of misinformation. Potential areas of extension might include investigating which regions produce more tweets with misinformation and the area's associated political partisanship to investigate the relationships between partisanship and misinformation spread [9].

Through these results and inferences, a workflow that extends off of this investigation may be useful to determine more granular insights into starting with topics or tokens that are most prevalent in misinformation tweets, subsetting the data based on topics of interest, and performing calculations to determine the sentiment associated with this subset of data, the location of these tweets, and finally, the number of retweets these tweets receive. This may benefit public health officials in soliciting which topics or issues may need public addressing or explanations to alleviate misinformation. It may also provide assistance in helping Twitter establish more targeted policies that reduce the spread of misinformation from verified accounts.

Conclusion

We built a machine learning model to determine specific properties that marked each tweet as misinformation or not misinformation and analyzed consequent reactions of vaccine hesitancy and compliance to health guidelines. From the data, we can observe an association between ver-

ified/unverified accounts and the spread of misinformation, as well as the public's willingness to heed safety policies based on social media consumption.

As more doses of vaccines are administered worldwide, researchers are continuously extracting social media data to filter out misinformation and protect the public against the risk of transmission and severe disease. Indeed, concern over misinformation in this digital age is growing and extant research seeks to unveil the effects and mechanisms by which it spreads. The predictive model built in this study lays the groundwork for such research and serves as a reliable tool for classifying information and offering insight into public sentiment. Such an understanding of people's emotions and behaviors will enable appropriate dissemination of health information as well as large-scale opportunities to educate the public.

Future work should investigate the cognitive psychological profile of individuals who fall victim to misinformation to better direct messages of public health measures and understand how socio-economic-demographic factors impact vulnerability.

Ultimately, the accuracy of this model facilitates a broader understanding of social media's influence on the general public in regards to infodemiology. Although the social dynamics between social media and content consumption is a complex research subject, it is vital in tailoring communication strategies during a pandemic and forecasting the spread of COVID-19 based on social perceptions and behavioral responses to public health interventions.

Acknowledgements

We would like to acknowledge Anish Verma and the team at STEM Fellowship for organizing this event. We would also like to thank our unofficial mentor, Srikar Katta, for reviewing our report.

References

[1] Fact sheet President biden to announce all americans to be eligible for vaccinations by may 1, puts the nation on a path to get closer to normal by july 4th. 2021.

[2] Coronavirus (covid-19) vaccinations - statistics and research. 2021.

[3] Mohammad S Razai, Tasnime Osama, Douglas GJ McKechnie, and Azeem Majeed. Covid-19 vaccine hesitancy among ethnic minority groups, 2021.

[4] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22:100104, 2021.

[5] K Hazel Kwon, J Hunter Priniski, and Monica Chadha. Disentangling user samples: A supervised machine learning approach to proxy-population mismatch in twitter research. *Communication Methods and Measures*, 12(2-3):216–237, 2018.

[6] Matthew R DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. *arXiv preprint arXiv:2101.07694*, 2021.

[7] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive aggressive algorithms. 2006.

[8] SV Praveen, Rajesh Ittamalla, and Gerard Deepak. Analyzing the attitude of indian citizens towards covid-19 vaccine–a text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(2):595–599, 2021.

[9] Sebastián Valenzuela, Daniel Halpern, James E Katz, and Juan Pablo Miranda. The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation. *Digital Journalism*, 7(6):802–823, 2019.