STEM Fellowship Big Data Challenge 2021

Evaluation of Twitter's Preventative Measures on the Spread of Misinformation on the COVID-19 Vaccine

Sina Haghighi[1], Glenn Chang[2], Kamakshi Naidu[3], and Harleen Kaur[1]

[1]Simon Fraser University
[2]University of Toronto
[3]University of Western Ontario

May 31, 2021

Abstract

Twitter has increasingly been the hub for people to obtain daily news and information, yet it is plagued with misinformation and fake news. During the coronavirus pandemic, information regarding the COVID-19 virus and health policies were constantly changing and it was difficult to differentiate between real or fake news. As such, this poses a serious health risk to society. Twitter has taken steps to combat this by utilizing machine learning to detect and delete misinformation, adding labels to misleading tweets and adding an exploration tab for accurate COVID-19 information. Our study aimed to investigate the effectiveness of these policies by looking at the frequency of COVID-19 vaccine misinformation before and after Twitter added these policies. As a proxy measure to quantify misinformation, we examined the tweet sentiment and subjectivity value, and found trends in common words and hashtags. In this preliminary study, we have found a general decreasing trend in misinformation after Twitter implemented these preventative measures suggesting that these policies are effective. Further study will need to confirm this as major assumptions in misinformation and sentiment value were made and various confounding variables may have affected the data. By studying Twitter's misinformation policy during COVID-19, we hope other social media companies will continue to combat misinformation.

Keywords
COVID-19, Vaccine, Twitter, Misinformation, Infodemiology, Data, Sentiment Analysis, Subjectivity Analysis, Preventative Measures

1 Introduction

The internet has become a significant hub of health-related data for people around the world. COVID-19 brought about by SARS-CoV-2, set off a pandemic quest for data with wide dispersal of false or misdirecting health data. False reports about looming food shortages provoked individuals to store supplies from the get-go in pestilence and caused real deficiencies. An individual has even died from ingesting chloroquine after reports referenced hydroxychloroquine as a potential treatment for Coronavirus.[1] The World Health Organization noted in 2019 that one of the greatest dangers to worldwide health was vaccine hesitancy.[2] Amidst the rise of COVID-19 vaccines, vaccine hesitancy and misinformation spread like wildfire. Misinformation and hesitance regarding vaccine efficacy, political or pharmaceutical doubt and legitimacy of COVID-19 were spread amongst society. With the ease of browsing, creating and sharing data on social media, the range and spread of misinformation has exponentially increased. During the course of this pandemic, social media has frequently been reprimanded for aiding in the spread of false news. Specifically, Twitter has been a common social media for COVID-19 misinformation and vaccine fear mongering with claims such as the Coronavirus immunizations are utilized to hurt or even destroy certain races.[3] Twitter aimed to combat this by utilizing machine learning algorithms to detect misinformation to flag and deleting dangerous false tweets about COVID-19. Misleading information may also be labeled with “"may contain misleading information about COVID-19 vaccines”.

Moreover, Twitter even permanently bans accounts that frequently post misinformation. The effectiveness of these measures for reducing misinformation have yet to be investigated.[4] Interestingly, we performed a preliminary test by tweeting out 30 tweets containing various blatant misinformation regarding COVID-19 vaccine in the tweet text and hashtag. Over the course of a week, it has not been deleted or flagged as misinformation. Thus, we believe that Twitter's preventative measures for misinformation should be investigated for their effectiveness. It is often quite difficult to measure misinformation, especially for COVID-19. New information regarding the virus was constantly being discovered during the pandemic. This study aims to use proxy measures to quantify misinformation in COVID-19 vaccines, sentimental value, subjectivity value, trends in hashtags and trends in common words. Sentimental analysis quantifies the general feeling and perspective of a certain sentence. Such methodologies are often used in study to analyze general perception. As this study only looks at COVID-19 vaccines tweets, tweets with negative sentiment and high subjectivity are more likely to be misinformation, thus in this study we use these measures as a proxy measurement for misinformation. Moreover, key hashtags and words that are found in tweets containing misinformation are also used as an indirect measurement for misinformation. Our study aims to evaluate the effectiveness of Twitter's misinformation algorithm and policies. This is done by investigating the change in misinformation in COVID-19 vaccines tweets overtime. We have found that there is a general decreasing trend in misinformation after the policy has been put in place.

2 Materials & Methods

2.1 Data Collection

Tweets about COVID-19 vaccines were collected from two databases, Kash's "Covid Vaccine Tweets" database[5] and Preda G. "All COVID-19 Vaccines Tweets" database[6]. Specifically, the tweet text, date and hashtags were collected. Tweets that were missing tweet text or date were excluded from the data. Kash's database collected trending tweets that included the hashtag "Covidvaccine" and out of the 207006 tweets in the database, 206993 were included in this study. Tweets from the Preda G. database are from searching various COVID-19 vaccine-related terms. All 78319 tweets from the Preda G. dataset were used in this study. A total of 28,5312 tweets were included in our study, ranging from January 2020 to May 2021.

2.2 Sentiment Analysis

As a proxy quantitative measurement of misinformation, a tweet sentiment value was measured. To quantify sentiment value, the Sentiment Intensity Analyzer method from the Natural Language Toolkit library was used. This quantified a tweet text sentiment from -1 as negative sentiment to +1 as positive sentiment. For sentimental analysis, the tweets were separated into positive, neutral and negative sentiment. Tweets with -1 to -0.33 sentimental value were considered to be negative, tweets from -0.33 to 0.33 were considered neutral and tweets from 0.33 to 1 were considered positive. An assumption was made that negative sentiment tweets about COVID-19 vaccines were likely misinformation. Thus, the sentimental value was tracked over time to observe changes in average sentiment values before and after the implementation of Twitter's misinformation policies.

2.3 Subjectivity Analysis

Similar to sentiment analysis, subjectivity was used as a proxy measurement for misinformation. The assumption that less subjective tweets were less likely to be misinformation. The subjectivity of a tweet text was measured using the subjectivity method in the TextBlob library. Subjectivity values are floats ranging from 0 to 1 where tweets with 0.0 subjectivity are considered very objective and 1.0 subjectivity is considered subjective. The monthly subjectivity values of tweets to be measured over time to determine changes before and after the implementation of various misinformation policies from Twitter. Specifically, changes in subjectivity in negative, neutral or positive tweets after implementation of these policies.

2.4 Hashtag Analysis

In order to decompose our problem, one of the fields that was a good correlation to false information is the use of hashtags in tweets. Setting time as our denominating factor, we collected a data set of 50,000 tweets[5] and analyzed the frequency of certain hashtags. Additionally, it includes a set of important events in the operation of Twitter's prevention methods, such as the blocking of profane or potentially misleading tweets, whose accuracy was iteratively improved is displayed in the plots. With the methods described, we correlated data from our sentimental value and subjectivity analysis and determined three of the most frequently used hashtags from both the negative and positive sentimental values. Over the period of 17 months, the data was plotted as a line plot with markers at each data point[7]. According to this theme, the date where significant events related to Twitter's prevention methods[4] were noted and marked on the plot as vertical lines.

Finally, we decided upon the following hashtags to represent:

Positive Tweets

- #CovidVaccine
- #Pfizer
- #stayhome

Negative Tweets

- #trump
- #wuhan
- #lockdown

2.5 Word Analysis

The packages 'tidyverse' and 'tidytext' in RStudio were used to tokenize words from Twitter tweets and stop words such as 'a' and 'the' were removed from the dataset along with common irrelevant words. We then analyzed 3 to 4 of the highest frequency of positive, neutral, and negative sentimental words over a span of 1 year. The sentimental value was also correlated with the frequency to confirm that the words being selected were associated with their sentiment. The results were then plotted over 'time' vs 'frequency' using the 'ggplot' package to determine whether specific words showed greater frequency around the time preventative measures were implemented.

3 Results

3.1 Sentiment Analysis

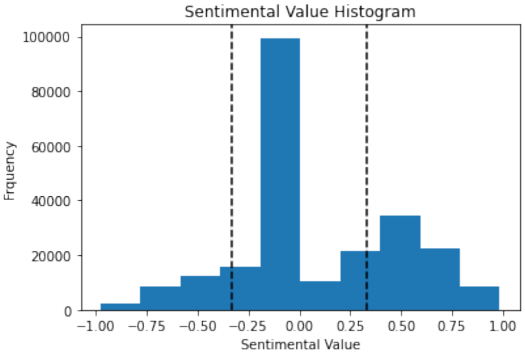

Figure 1: The spread of sentimental values across the whole dataset. The dotted lines represent the threshold between negative, neutral and positive sentiment. Tweet sentiment calculated using NLTK library.

First, the sentiment of each tweet was calculated. Figure 1 shows the spread of sentiment values in our whole data set. The majority of the tweets were considered neutral sentiments. Specifically, 161,392 tweets were considered neutral, 89.851 tweets were positive and only 34,069 tweets were negative. Under our assumption that negative sentiment is correlated with misinformation, most tweets in our data set were not considered misinformation.

Twitter implemented various misinformation policies and algorithms to combat COVID-19 misinformation during the course of the pandemic. The main changes include implementing a machine-learning algorithm to detect and delete misinformation tweets on April 1st, 2020. May 18th, 2020, for adding misinformation labels to inaccurate COVID-19 tweets. Lastly, May 11th, 2020 for adding search prompt and exploration tab that provides accurate up-to-date information regarding COVID-19 [4]. These dates will be used as thresholds to determine if there is a change in sentiment and subjectivity values.

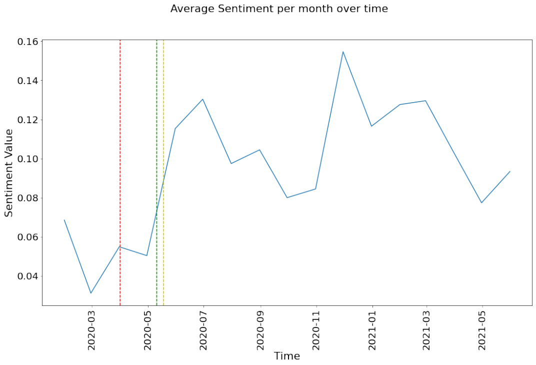

Figure 2: The average monthly sentiment over time. Tweet sentiment calculated using NLTK library. The red, green and yellow dotted line represent dates twitter implemented misinformation detection algorithm, misinformation labels, accurate COVID-19 exploration tab respectively

Figure 2 shows the average monthly sentiment of the whole dataset over time. The tweets were grouped by month and the average sentiment is calculated for each month. The key dates for implementation of twitter misinformation policies are April 1st, 2020 for implementing machine learning algorithms to detect misinformation for deletion, May 11th, 2020 for implementation of misinformation labels under tweets that contain COVID-19 misinformation and May 18th, 2020 for implementation of COVID-19 explore tab and search prompts for easier access to accu-

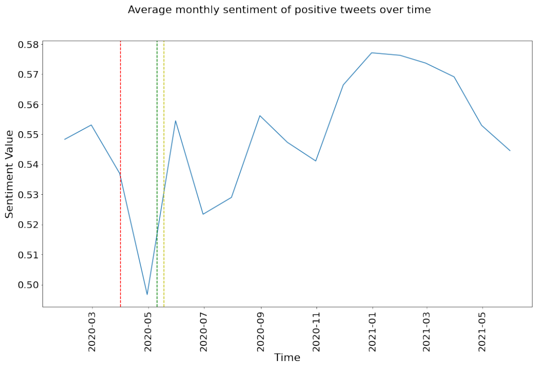

Figure 3: The average monthly sentiment of positive tweets over the course of the pandemic. Tweet sentiment calculated using NLTK library. The red, green and yellow dotted line represent dates twitter implemented misinformation detection algorithm, misinformation labels, accurate COVID-19 exploration tab respectively

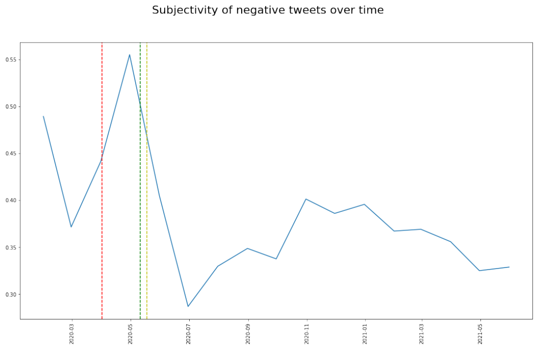

Figure 4: The average monthly sentiment of negative tweets over the course of the pandemic. Tweet sentiment calculated using NLTK library. The red, green and yellow dotted line represent dates twitter implemented misinformation detection algorithm, misinformation labels, accurate COVID-19 exploration tab respectively

rate information regarding COVID-19 misinformation. As shown in figure 2, the average sentiment values before the implementation of these policies are much lower, ranging from 0.03 to 0.06 from February 2020 to May 2020. After the implementation, the average monthly sentiment value increased substantially, ranging from 0.08 to 0.15 from June 2020 to June 2021. As the average monthly sentiment increased after these policies, this suggests that there was a decrease in misinformation tweets about vaccines.

Figures 3 and 4 show the change in average monthly sentiment values in the positive tweets and negative tweets separately. Figure 3 shows the change in sentiment values of positive tweets over time. There was a major decrease in sentiment value of 0.49 during May 2020, this was likely due to a sharp increase in COVID cases and prolonged lockdown in North America. However, other than May 2020, overall, there were very few changes in sentiment values before and after the implementation of the three Twitter policies described above. As we assumed only negative sentiment tweets were misinformation and positive sentiment tweets were not, these misinformation policies should not affect the positive tweets thus as illustrated in figure 3A, there were no major changes in sentiment values after misinformation policies were put in place.

In contrast, Figure 4 shows the change in average monthly sentiment values in negative tweets over time. There is a clear difference in sentiment values before and after twitter put these misinformation policies into effect. Before the misinformation policies, the sentiment values ranged from -0.61 to -0.53 from February 2020 to May 2020. However, after policy implementation, the average monthly sentiment values were higher, ranging from -0.56 to -0.47 from June 2020 to June 2021. This suggests that there was less misinformation after the implementation of twitter misinformation policies. This is in stark contrast with positive tweets in figure 3 which showed no major changes before and after implementation of the policies.

3.2 Subjectivity Analysis

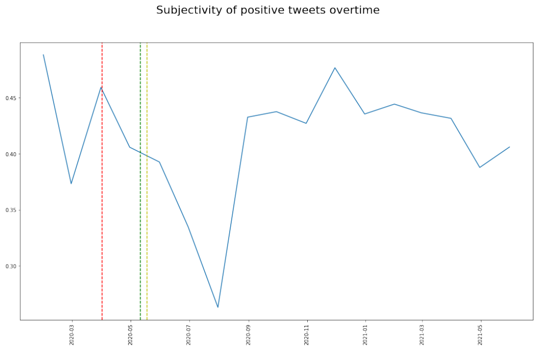

Figure 5: The average monthly subjectivity over time. Subjectivity of tweets were calculated using the Textblob library. The red, green and yellow dotted line represent dates twitter implemented misinformation detection algorithm, misinformation labels, accurate COVID-19 exploration tab respectively

Figure 5 shows the average monthly subjectivity of all tweets over time. Tweets were grouped by month and the average subjectivity was calculated. Before twitter's actions against COVID-19 misinformation, there was a slightly higher subjectivity value with the peak of 0.36 during April 2020. However, after policy was put into place, there was a sharp drop in subjectivity of 0.26 and 0.24 during July and August of 2020 respectively. This may be because tweets that were subjective were classified as misinformation and deleted. Interestingly, the subjectivity increased in the subsequent months. Although it is unclear what caused this increase in subjectivity, this could be due to people's frustration over COVID-19 or excitement regarding success in COVID-19 vaccine studies. Another possible explanation is that Twitter also adjusted their classification of misinformation during July 2020 [4] . The overall decreasing trend in average monthly subjectivity after twitter policy implementation suggests that there is a decreased misinformation.

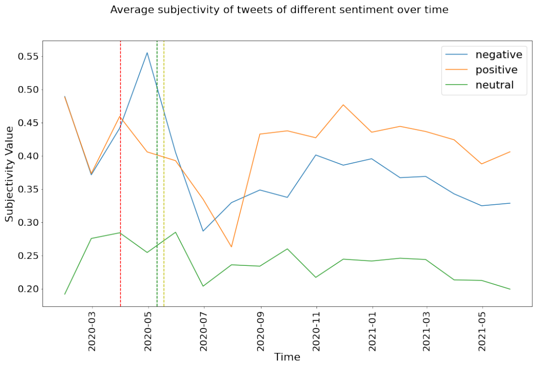

Figure 6: The average monthly subjectivity for positive, negative and neutral tweets over time. Subjectivity of tweets were calculated using the Textblob library. The red, green and yellow dotted line represent dates twitter implemented misinformation detection algorithm, misinformation labels, accurate COVID-19 exploration tab respectively.

Subjectivity is not a direct measure of misinformation as positive tweets about COVID-19 vaccines will also have a higher subjectivity value. Hence, tweet subjectivity with positive, neutral or negative sentiment is investigated separately. Figure 6 illustrates the change in average monthly subjectivity for different sentiment tweet groups. The green line shows the subjectivity of neutral tweets. As expected, the subjectivity of neutral tweets is relatively low compared to positive and negative tweet sentiment as neutral tweets are likely more objective. The orange line illustrates the average subjectivity of positive sentiment tweets. Other than the large decrease in subjectivity value during July 2020, the subjectivity value remains the same over the course of the pandemic. The policy implemented by Twitter had no major effect on subjectivity for positive sentiment tweets as positive sentiment tweets are unlikely to be misinformation. On the other hand, negative sentiment tweets, represented by the blue line, show a drop after Twitter introduced its misinformation policies. Before the policies, negative tweets had high subjectivity values indicating that tweets were very subjective. These values range from 0.37 to 0.55 from February 2020 to May 2020. However, after the policies, negative tweets had an overall lower subjectivity value ranging from 0.28 to 0.40 indicating the negative tweets were more objective. The decreasing trend in subjectivity found only in negative sentiment tweets but not in positive and neutral sentiment tweets after misinformation policies were implemented suggests that there were fewer misinformation tweets due to the policy changes.

3.3 Hashtag Analysis

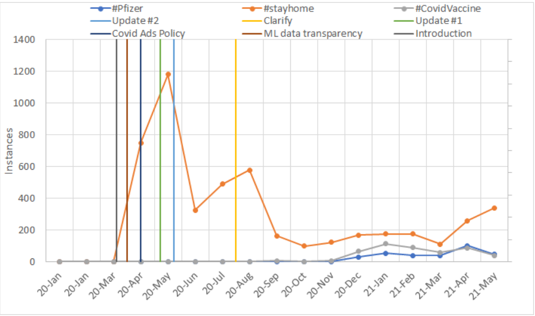

Figure 7: Demonstrates the trend of 3 common hashtags that correlated to positive and neutral sentimental tweets spanning over time, including the updates of Twitter's preventative measures.

In Figure 7 we observe the results from our positive and neutral sentimental tweets. There is a notable dispersion of the tweets and their frequency of use over time. Notice the new updates to the Twitter prevention methods (presented by vertical lines), have little to do with surge in #stayhome tweets as they rise. This signifies that Twitter is not preventing the spread of positive sentimental tweets.

Figure 8 demonstrates the results from the most frequently used tweets in the negative sentimental tweets. According to the results, the preventative measures are having a greater impact on the spread of negative tweets. Notably with there is a direct correlation between the steady roll-out of updates decreased the uses of

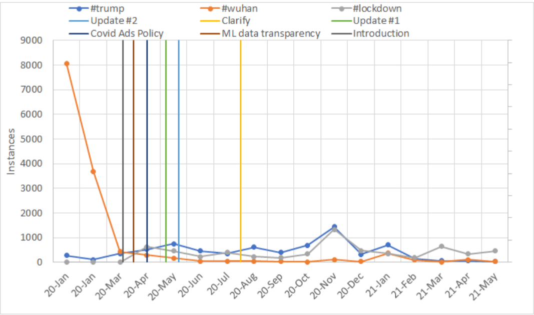

Figure 8: Demonstrates the trend of 3 common hashtags that correlated to negative sentimental tweets spanning over time, including the updates of Twitter's preventative measures.

negative sentimental-related hashtags, namely #lockdown and #Wuhan.

3.4 Word Analysis

The highest frequency words for positive sentiment were "moderna" (54.7%), "effective" (20.9%), and "pfizer" (24.2%) out of 685,748 words (Figure 9). The highest frequency words for neutral sentiment were "covishield" (15.0%), "dose" (63.2%), and "health" (21.6%) out of 1,164,701 words (Figure 10). The highest frequency words for negative sentiment were "astrazeneca" (22.2%), "forced" (23.0%), and "death" (23.9%) out of 1290,819 words (Figure 11). Neither positive, neutral, or negative words showed a significant change in frequency when preventative measures were implemented.

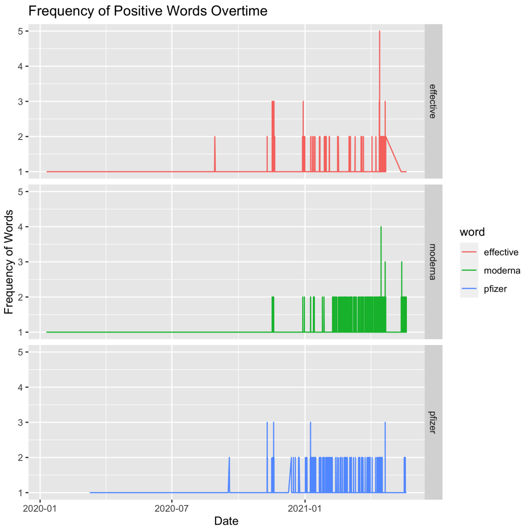

Figure 9: Demonstrates the frequency trend results of the 3 highest frequency positive words over a span of year

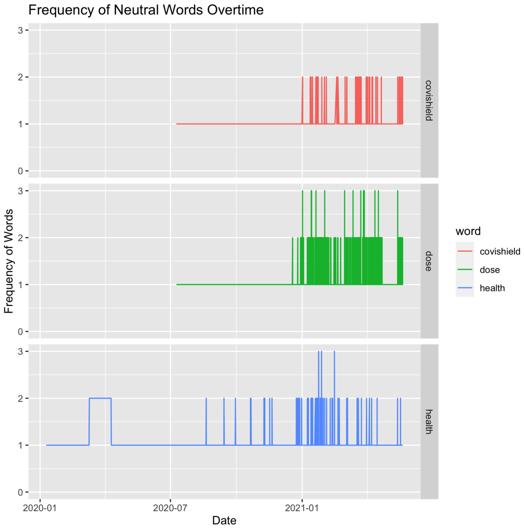

Figure 10: Demonstrates the frequency trend results of the 3 highest frequency neutral words over a span of year

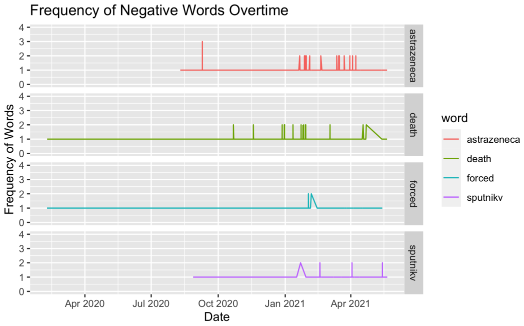

Figure 11: Demonstrates the frequency trend results of the 4 highest frequency negative words over a span of year

4 Discussion

This study aims to evaluate the effectiveness of Twitter's preventative measure against the spread of misinformation. One method of indirectly measuring misinformation is sentiment values and subjectivity values. We assume that tweets about COVID-19 vaccines with negative sentiment and high subjectivity are likely misinformation. Our data only focuses on COVID-19 vaccine tweets as many fear-mongering tweets about the vaccines were being spread to dissuade people from getting the vaccine. Twitter considers tweets to be misinformative based on tweets that could lead to harm or influence user's behaviours and beliefs regarding claims that are disproven. As these tweets are likely negative and subjective, our study hypothesizes that these tweets with those characteristics should decrease after the implementation of these policies. By comparing positive and neutral tweets with negative tweets, the effectiveness of twitter's misinformation policy can be evaluated as it should only affect negative tweets. This is shown in figure 3, where the average monthly sentiment of positive tweets remains relatively constant even after twitter's new policies because they likely will not be flagged as misinformation.

Contrastingly, negative sentiment tweets showed an increasing trend for sentiment after the policy was implemented which could be due to twitter deleting or flagging them as misinformation. These results are further supported by the subjectivity values. As shown in figure 5, positive and neutral tweets had relatively constant subjectivity values and were not affected by the policies from Twitter. On the other hand, negative tweets showed a decrease in subjectivity values which could be due to very subjective misinformation tweets being deleted. Collectively, the subjectivity and sentiment values point to twitter's misinformation policies being effective at reducing misinformation being spread on Twitter. However, further statistical analysis is needed to confirm if the changes are statistically significant or not.

In terms of the hashtags, our study has found that negative hashtags associated with misinformation were found in lower frequencies after Twitter's policies. As an example, #stayathome and lockdown have similar usage however, #stayathome has a more positive connotation whereas lockdown has a more negative connotation, thus it is likely that misinformation tweets used the #lockdown more than #stayathome. We see that the frequency of lockdown decreased after Twitter policies were implemented which could be due to the deletion of these misinformation tweets whereas #stayathome showed a similar frequency during the whole pandemic. However, based on the hashtag analysis, these policies were not always enough to reduce the spread of misinformation, especially when it relates to certain real-world events. For example, the US presidential election in the Fall of 2020 caused a spike in misinformation tweets using the word trump which was not affected by Twitter's policies. Based on the data presented there is a relationship between the preventative measure updates that limited the use of negative sentimental hashtags but not positive and neutral sentiment hashtags.

Contrastingly, the frequency of specific words with negative sentimental (Figure 11) value did not change after the preventative measures were implemented. This suggests that these preventative measures do not impact the spread of misinformation on Twitter. However, this can be due to the small sample size used in our data, we had a limited amount of data for negative words. We were not able to find a common word that is consistently found only in tweets with misinformation. As the negative words are not a strong representation of misinformation, the frequency was largely stable throughout late 2020 and early 2021.

A few limitations within our study, mainly, our study assumes that misinformation tweets likely have negative sentiment and high subjectivity. Although the study only used COVID-19 vaccine tweets to reduce this limitation, this is not a completely accurate assumption. Tweets that are pointing out valid concerns about long-term health outcomes of the COVID-19 vaccine or general frustration toward vaccine roll-out schedule would potentially be classified as negative and subjective. As this was a preliminary study to show the general trend of misinformation after twitter's policy, a future study would use an improved measure of misinformation to determine the policy's effectiveness. For example, using a neural network to classify tweets as real or fake and determining the frequency of real or fake tweets before and after twitter's misinformation policy. This would allow for a better indicator of misinformation however, due to time constraints, our study could only use sentiment and subjectivity as a proxy for misinformation. Moreover, many other factors also affect sentiment values and hashtag usage. For example, as the vaccination rate begins to increase, there would be an increasing positive sentiment compared to earlier during the quarantine. In terms of hashtag usage, the Trump hashtag may also be affected by the US presidential election this is not representative of misinformation. Further investigation into other potential compounding variables should be done to ensure that there truly is a change in misinformation.

Another limitation is the low number of tweets in our data set. Our data set contains 285,312 tweets, however since this study specifically looked at vaccines, most of the tweets were in early 2021. The small number of tweets in early 2020, before Twitter implemented their misinformation policies, may have resulted in less accurate sentiment and subjectivity values. This may be a potential factor affecting the difference in sentiment, subjectivity, and word frequency before and after the policy. Other potential studies can look at topics that would not have a bias in the number of tweets for a specific duration.

Conclusion

The purpose of this study was to evaluate the efficacy of Twitter's preventative measures on the spread of misinformation about the COVID-19 Vaccine. It was found that the measures put in place by Twitter demonstrate effectiveness for the intended purpose of deleting or flagging misleading tweets. Socioeconomic events play a significant role in the fluctuation of misinformation and override efforts set in place by Twitter's current preventative measures. Currently, Twitter's system can recognize significant misleading information by analysing the behaviour of the accounts with odd behaviour and block it but is not able to perform further analysis to capture a broader scope of tweets that are being missed with its latest implementation[**?**].

All in all, our study has found that based on sentiment value, subjectivity and hashtag usage of COVID-19 tweets, there is a decrease in misinformation after Twitter has implemented changes to their misinformation policy. This suggests that these policies are effective at combating the spread of misinformation. As this is just a preliminary study, we hope to further investigate the effectiveness of each of these policies and improve the quantification of misinformation. With the broader technology space becoming more vigilant with the type of information that is spread, companies such as Google and Facebook are investing millions of dollars into information integrity, it may be time for other social media giants to join too.

Acknowledgements

We would like to thank our mentors Shan Dhaliwal, Bryce Haley and Anish Verma for their guidance and recommendations for this project. More broadly, we would like to thank STEM Fellowship, for hosting and facilitating this event for students to engage in data science.

References

[1] WHO. Immunizing the public against misinformation, Aug 2020.

[2] Griffith J;Marani H;Monkman H;. Covid-19 vaccine hesitancy in canada: Content analysis of tweets using the theoretical domains framework, Apr 2021.

[3] Meeyoung Cha1, Chiyoung Cha3, Karandeep Singh2, Gabriel Lima1, Yong-Yeol Ahn4, Juhi Kulshrestha7, Onur Varol8, 1School of Computing, and Corresponding Author:Meeyoung Cha. Prevalence of misinformation and factchecks on the covid-19 pandemic in 35 countries: Observational infodemiology study, Feb 2021.

[4] Twitter Inc. Coronavirus: Staying safe and informed on twitter, Jan 2021.

[5] Kash. Covid vaccine tweets, May 2021.

[6] Gabriel Preda. All covid-19 vaccines tweets, May 2021.

[7] Christian Lopez and Malolan Vasu. lopezbec/covid19$_{t}weets_{d}ataset, Feb$2021.