2020 7th International Conference on Behavioural and Social Computing (BESC) | 978-1-7281-8605-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/BESC51023.2020.9348291

Sentiment, Count and Cases: Analysis of Twitter discussions during COVID-19 Pandemic

Zainab Tariq Soomro
Lahore University of Management Sciences
Email: zainabsoomro@gmail.com

Sardar Haider Waseem Ilyas
Lahore University of Management Sciences
Email: haider@waseemilyas.com
Web: www.haiderwaseem.com

Ussama Yaqub
Lahore University of Management Sciences
Email: ussama.yaqub@lums.edu.pk

***Abstract*—In this paper, we analyze over 18 million coronavirus related Twitter messages collected between March 1, 2020 and May 31, 2020. We perform sentiment analysis using VADER, a rule-based supervised machine learning model, to evaluate the relationship between public sentiment and number of COVID-19 cases. We also look at the frequency of mentions of a country in tweets and the rise in its' daily number of COVID-19 cases. Some of our findings include the discovery of a correlation between the number of tweets mentioning Italy, USA, and UK and the daily increase in new COVID-19 cases in these countries.**

Keywords — COVID-19, Sentiment Analysis, Twitter

1. Introduction

In December 2019, unknown pneumonia cases were first reported by the Chinese government. Since then the pandemic has spread globally, which is named as COVID-19. As the disease spread, people worldwide started using social media extensively to express their opinions regarding COVID-19. People used Twitter - a popular micro-blogging platform - to express their views related to COVID-19, such as the idea of that coronavirus is somehow related with 5G roll-out or other misinformation related to the virus' spread and cure [1], [2].

This rich information has allowed researchers to utilize Twitter data to analyze the pandemic. There have been studies looking into tweets of different world leaders and their messages to the public since the start of the pandemic [3].

In this paper, we evaluate the sentiment of tweets gathered for 92 days - starting from Mar 1, 2020 to May 31, 2020 - mentioning the terms 'corona' or 'coronavirus'. We utilize the Python VADER library to perform sentiment analysis. We matched the sentiment with the rise in cases in different countries that were hit hard by the pandemic during this period. We make the following contributions:

- Comparison of the sentiment of tweets mentioning Italy, United States, and United Kingdom with the rise in COVID-19 cases.
- Comparison of the number of tweets mentioning Italy, United States, and United Kingdom, with the number of COVID-19 cases in the countries.

In the next section, we review some of the previous works related to sentiment analysis of Twitter. In sections 3 and 4, we discuss our methodology and perform data analysis, respectively. Section 5 contains a discussion of the results before we finally conclude the paper in section 6.

2. Previous Work

With the rise in popularity of social media, sentiment analysis of online user discussions has become a very active research area. Twitter is one of the most popular social networks having over 350 million monthly active users [4]. This massive user base has made it an attractive source of data collection for analysis [5], [6]. Thus, from predicting approval of electoral candidates to gauging the popularity of anti-government protests, applications of sentiment analysis on tweets have been far and wide in academic research [7], [8]. Sentiment analysis of tweets during elections has especially seen substantial growth as a research topic [9], [10].

The COVID-19 pandemic has been a popular topic of discussion in online social media platforms. There have been researches on the spread of fake news and conspiracy theories [1], [2]. Studies have also looked at the spread of misinformation on different social media platforms. Cinelli et al. looked at the spread of COVID-19 related information diffusion across multiple social media platforms such as Twitter, Instagram, YouTube, Reddit, and Gab [11]. The study discovered Gab as having an environment more susceptible to the spread of misinformation.

With the increase in the application of sentiment analysis on Twitter data, different tools and techniques needed to perform nuanced analysis have also evolved [12], [13]. Recently VADER has become a popular model for sentiment analysis [13]. The latest studies performing sentiment analysis of Twitter data have utilized VADER to gauge user sentiment [6], [14]. The tool has been claimed to perform as well as humans in gauging the sentiment of social media messages [13].

3. Methodology

In this section, we discuss the key features of our methodology.

3.1. Data Collection

For this study, we used Twitter search API to collect tweets pertaining to the COVID-19 pandemic over a period of 92 days, ranging from Mar 1, 2020 to May 31, 2020 [15]. This resulted in a total sample size of over 18 million tweets. For the purpose of our analysis, we only extracted the date and text variables. Furthermore, we gathered the daily increase in COVID-19 cases in Italy, USA, and UK for the same period.

3.2. Word Cloud

The first step to generate the word cloud involved tokenizing the tweets in our data set. Tokenization is the process of breaking down the tweets into words. The words were then converted to lowercase to ensure that words such as ’Day’ and ’day’ are considered the same. This was followed by removing stop-words to help filter out words that do not add meaning to the discussion. Finally, we counted the number of times each word appeared in our data set and passed the counts to Python’s word cloud library in order to generate the plot.

3.3. Sentiment Analysis

We performed sentiment analysis on the tweets in their original form - without any pre-processing - in order to capture their true sentiment. For this task, we utilized Python’s VADER library, a lexicon and rule-based sentiment analysis tool specifically tailored to sentiments expressed in social media [13]. VADER’s major advantage is that it does not simply categorize text into sentiments such as positive, negative, or neutral. Instead, it gives the proportion of the text that falls into each category. Furthermore, it gives a compound score that assigns a polarity to the overall text. This score ranges on a scale of -1 to +1, with -1 being the most negative and +1 being the most positive. [16]. Our analysis focuses on the compound scores of the tweets.

3.4. Country-wise Counts, Sentiment and Cases

For this section, we analyzed our tweets with respect to the countries they mentioned. The countries we chose were Italy, USA, and UK as they were most affected by COVID-19 during the time period of our data set. We extracted three subsets of data, one for each of the countries mentioned above. On each of the three data sets, we calculated the daily mean sentiment and the daily number of tweets.

We compared the daily mean sentiment and the daily number of tweets for each country to the daily increase in the COVID-19 cases for that country. To test the relationships between these variables, we employed the Spearman Rank Correlation. For plotting purposes, we normalized the values between 0 and 1.

TABLE 1: Number of Tweets collected for Analysis

Total tweets	18,172,848
Mean daily tweets	197,530
Percentage unique tweets	34.56%

Figure 1: Wordcloud containing frequently used words.

4. Data Analysis

In this section, we discuss the main aspects of our analysis.

4.1. Wordcloud

From figure 1, we can observe that the name ’Trump’ has been mentioned frequently in the discussions. Furthermore, the discussions predominately contain various COVID-19 related terms such as ’virus’, ’pandemic’, ’health’, ’lock down’, ’cases’, and ’deaths’. As expected, the mention of China is also prevalent in the discussion.

4.2. Sentiment Analysis

For a majority of days in our data set, the histogram in figure 2 shows a sentiment below zero. This indicates that the Twitter discussions revolving around COVID-19 had a negative connotation attached to them. It is also clear that despite the fluctuations, the daily mean sentiment is negative for all days except two. The most significant drop is experienced around Mar 7, 2020 which corresponds to the time when the global number of COVID-19 cases rose above 100,000 and the total number of countries having reported confirmed COVID-19 cases crossed 100. Moreover, it was on Mar 7 that the Italian cabinet started discussing lock down, which was subsequently implemented on Mar 8. Another significant drop is observed on the last day of our data set - May 31, 2020 - which is when the global COVID-19 cases count surpassed 6 million. It was also around that time when the death toll in USA, due to COVID-19, crossed the 100,000 mark.

4.3. Country-wise Sentiment, Count and Cases

In this subsection, we will discuss the counts, sentiment and cases of each of the three countries: Italy, USA, and UK.

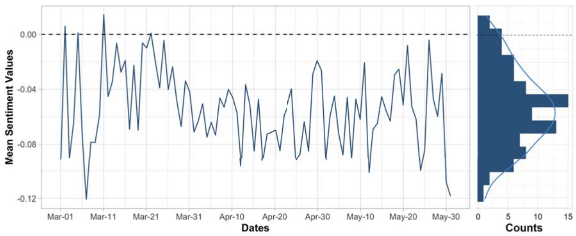

Figure 2: Daily mean sentiment scores and their distribution from Mar 1, 2020 to May 31, 2020.

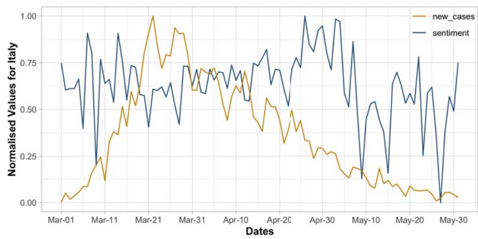

Figure 3: Time series plot of mean sentiment scores for tweets mentioning Italy and new cases in Italy

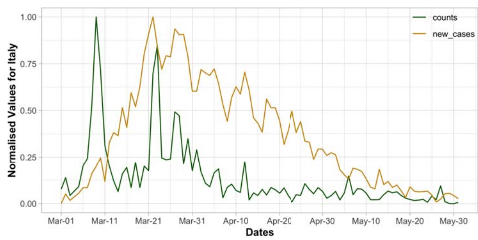

Figure 4: Time series plot of number of tweets mentioning Italy and new cases in Italy

4.3.1. Italy. Figure 3 displays a normalized time series comparison of the daily mean sentiment for tweets mentioning Italy and the daily increase in cases in Italy. We can see that the values of the two variables tend to follow a similar pattern on a few days. However, on most days, the two variables are not in sync with each other. To test this, we calculated the Spearman Rank Correlation coefficient. The result was 0.153, with a statistically insignificant p-value of 0.146, which shows that the two have an extremely weak linear relationship.

Next, we compare the daily number of tweets mentioning Italy with the daily increase in cases there. Through figure 4, we can see that the normalized values of these variables tend to follow a similar pattern. On several instances, the increase in the number of cases is accompanied by an increase in tweets. The correlation coefficient for the two variables came out to be 0.545, with a p-value of 1.94e-08, showing a statistically significant linear relationship.

4.3.2. USA. We can see from figure 5 that the daily mean sentiment for tweets mentioning USA and the daily increase in the cases in USA are not in sync with each other. These

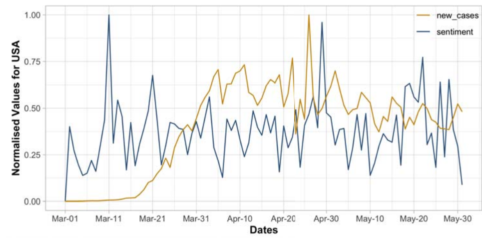

Figure 5: Time series plot of mean sentiment scores for tweets mentioning USA and new cases in USA

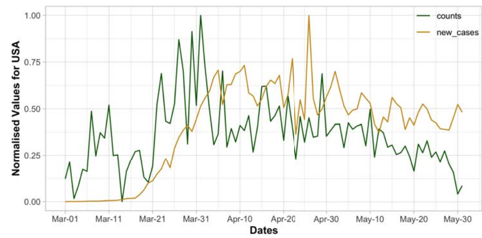

Figure 6: Time series plot of number of tweets mentioning USA and new cases in USA

results are similar to what we observed for Italy. When tested for correlation, the Spearman coefficient came out to be 0.098, with a p-value of 0.353. Thus, we can conclude that the two have an extremely weak linear relationship.

On the other hand, as seen in figure 6, the spikes in daily increase in cases in USA coincide with an increase in tweets mentioning USA. The Spearman correlation test of these variables resulted in a correlation coefficient of 0.415 and a p-value of 3.83e-05. Similar to Italy, this results in a statistically significant linear relationship.

4.3.3. UK. From figure 7, we can observe a normalized time series plot of the daily mean sentiment for tweets mentioning UK and the daily increase in cases in the UK. It is clear that the two variables do not tend to move together. This can be confirmed from the Spearman coefficient, which came out to be -0.127 with a p-value of 0.228, showing that no statistically significant linear relationship exists.

Figure 8 shows the daily number of tweets mentioning UK and the daily number of new cases in the UK. We can see that the two variables seem to move together. The Spearman correlation test of the two resulted in a coefficient of 0.485 and a p-value of 9.82e-07. This shows that there exists a statistically significant linear relationship . Both these results are similar to those of Italy and USA.

5. Discussion

In this study, we analyzed public discussions about COVID-19. We observe that the discussions include various terms related to COVID-19, such as virus and pandemic, from the word cloud. The words China and Trump are also frequently mentioned, reflecting the public perception of their involvement and actions related to the pandemic.

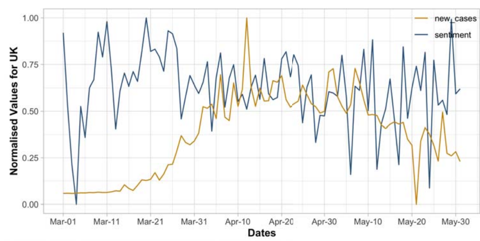

Figure 7: Time series plot of mean sentiment scores for tweets and new cases in UK

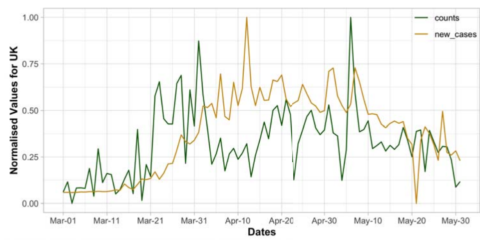

Figure 8: Time series plot of number of tweets mentioning UK and new cases in UK

Though this is not an exhaustive approach, it gives a macro perspective regarding the potential topics of the public's discussion.

Moreover, we arrived at an understanding of the public sentiment towards COVID-19. For the majority of days, the mean sentiment was negative, which corresponds to the grim nature of the pandemic. The negative sentiment was especially apparent during two days. First, around Mar 7, 2020, when the worldwide COVID-19 cases rose above 100,000 and the number of COVID-19 affected countries crossed 100. Second, around May 31, 2020, when the worldwide COVID-19 cases rose above 6 million while the number of COVID-19 related deaths rose above 100,000 in the USA. It goes to show that the sentiment of the Twitter discussion related to COVID-19 does incorporate actual events taking place at the time.

Finally, we analyzed the tweets of three countries - Italy, USA, and UK - with respect to their COVID-19 situation. The sentiment scores and daily increase in cases of COVID-19 in these countries did move together in a few instances. However, no statistically significant correlation existed between them. On the other hand, we observed a statistically significant correlation between the number of tweets mentioning the affected country and the daily increase in COVID-19 cases there, for all three countries. Thus, we can say that the public directs more attention to a country as the daily new cases increase.

6. Conclusion

In this paper, we analyzed over 18 million COVID-19 related tweets collected over a period of 92 days. We evaluated their sentiment and the number of tweets mentioning countries affected by the pandemic.

In the future, we would like to expand this study by performing topic modeling on the Twitter discussion. We would also like to evaluate the tweets' content for terms related to fake news and conspiracy theories.

References

[1] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, "Covid-19 and the 5g conspiracy theory: social network analysis of twitter data," *Journal of Medical Internet Research*, vol. 22, no. 5, p. e19458, 2020.

[2] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter," *Cureus*, vol. 12, no. 3, 2020.

[3] S. R. Rufai and C. Bunce, "World leaders' usage of twitter in response to the covid-19 pandemic: a content analysis," *Journal of Public Health*, 2020.

[4] "Twitter stats," https://www.statista.com/statistics/282087/number-of-monthly-active-twitter- users/, accessed: 2017-05-09.

[5] U. Yaqub, N. Sharma, R. Pabreja, S. A. Chun, V. Atluri, and J. Vaidya, "Location-based sentiment analyses and visualization of twitter election data," *Digital Government: Research and Practice*, vol. 1, no. 2, pp. 1–19, 2020.

[6] T. Mustaqim, K. Umam, and M. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," in *Journal of Physics: Conference Series*, vol. 1567, no. 3. IOP Publishing, 2020, p. 032024.

[7] U. Yaqub, S. Chun, V. Atluri, and J. Vaidya, "Sentiment based analysis of tweets during the us presidential elections," in *Proceedings of the 18th Annual International Conference on Digital Government Research*. ACM, 2017, pp. 1–10.

[8] A. Morales, J. Losada, and R. Benito, "Users structure and behavior on an online social network during a political protest," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 21, pp. 5244–5253, 2012.

[9] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, "Analysis of political discourse on twitter in the context of the 2016 us presidential elections," *Government Information Quarterly*, vol. 34, no. 4, pp. 613–626, 2017.

[10] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, no. 1, pp. 178–185, 2010.

[11] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *arXiv preprint arXiv:2003.05004*, 2020.

[12] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *Journal of the Association for Information Science and Technology*, vol. 62, no. 2, pp. 406–418, 2011.

[13] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[14] S. H. W. Ilyas, Z. T. Soomro, A. Anwar, H. Shahzad, and U. Yaqub, "Analyzing brexit's impact using sentiment analysis and topic modeling on twitter discussion," in *The 21st Annual International Conference on Digital Government Research*, 2020, pp. 1–6.

[15] T. API. (2018) Streamingapi. Accessed: 2018-04-01. [Online]. Available: https://dev.twitter.com/streaming/overview

[16] P. Pandey, "Simplifying sentiment analysis using vader in python (on social media text)," Nov 2019. [Online]. Available: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f