Opinions Analysis on Covid-19 Pandemic

Chen Meng
A0148018H
National University of Singapore
e0012690@u.nus.edu

Gu Zhehao
A0251091N
National University of Singapore
e0950095@u.nus.edu

Zhang Youyang
A0251290M
National University of Singapore
zhang.youyang@u.nus.edu

Zhang Zhirui
A0147973N
National University of Singapore
e0012645@u.nus.edu

Zhao Tian Qi
A0187506X
National University of Singapore
e0323090@u.nus.edu

ABSTRACT

Coronavirus disease (COVID-19) has spread to all continents since its first appearance. With the increasing amount of information shared on the internet, monitoring individuals' responses to health-related issues on social media has reflected useful information to study diseases. This project aims to analyze opinions and sentiments of relevant and pre-processed tweets in the period of October 2019 until December 2020, and find their impacts on society. The data is obtained from an open source and adjusted to fit this project. Analysis is applied in 3 dimensions: temporal analysis, influencer analysis, and frequent itemset analysis. The results of the study conclude that negativity increases in tweets after the announcement of COVID-19 outbreak by World Health Organization (WHO), entities in the same frequent itemsets share similar pattern in number of tweets and sentiment scores, and influencers show different behaviors in tweets than other groups of people due to their high public exposure. An ARIMA model is built to predict sentiment in tweets for future pandemic analysis.

KEYWORDS

COVID-19, Twitter, Opinions Analysis, Influencer Analysis, Sentiment Analysis

1 Background

COVID-19 has been affecting people's lifestyles such as remote working and social distancing since the World Health Organization (WHO) announced its outbreak in March 2020. With increasing COVID-19 related information shared online, the public shows high attention on this topic. Researchers should not only focus on people's physical health, but also their mental health through this long fight against COVID-19. Twitter, as one of the biggest social medias worldwide, provides a platform for the public to express their thoughts and sentiments. Through analysis of the public's opinions online, scientific models could be built to fit the public's sentiment trend or predict how government policies could affect the public's opinions, which would provide a powerful tool for future use in pandemic analysis.

Social media like Twitter is full of data and is one of the richest sources of real-world data, which also leads to the difficulty of crawling data from it. There are several Python packages that allow users to get data from Twitter such as Tweepy and Snscrape, but they have limitations either on data size or data functionalities. The analysis covered in this project requires a large size of data, so ready-to-use open source data becomes the optimal solution.

There already exists research that covers similar content in this project. For example, the website sharing the data completed a series of descriptive analysis, but it lacks predictive analysis that could be used for future pandemics. Several papers were found focusing on detailed analysis in a specific country such as Singapore. This project aims to provide more perspectives in this research area such as model fitting for sentiment trend and worldwide data analysis.

2 Data Source

Data is collected from a public website called TweetsKB, which contains a large amount of anonymous tweets. The website has more than 2 billion tweets data from Feb 2013 to Dec 2020. TweetsCov19 is a subset of TweetsKB which captures online discourse about various aspects of the pandemic and social impact. TweetsCov19 consists of

20,112,480 tweets posted by 7,384,417 Twitter users, spanning from Oct 2019 to Dec 2020.

The raw data is stored as tab-separated values (tsv) format and includes 12 features listed below:

1. Tweet id
2. Username: all usernames are encrypted for privacy issue
3. Timestamp: format "EEE MMM dd HH:mm:ss Z yyyy"
4. Number of followers
5. Number of friends
6. Number of retweets
7. Number of favorites
8. Entities: original text are aggregated for privacy issue and only annotated entities are displayed
9. Sentiment: positive and negative sentiment scores are produced by SentiStrength. The score range of positive and negative are [1,5] and [-1,-5] respectively.
10. Mentions: '@' are removed. If there are no mentions, the value is null.
11. Hashtags: '#' are removed. If there are no hashtags, the value is null.
12. URLs: URLs contained in the tweet if any.

3 Data Preparation

The raw data in tsv format has the size of 4.58G. The following preprocessing steps are done to clean and format data:

1. Dropped bad entries: some entries have missing columns or redundant columns that are not identifiable, therefore, these entries are removed.
2. Parsed entities and sentiment score: processed entity column to generate lists of entities that each tweet contains. Separated sentiment score into positive score and negative score and computed overall sentiment score by summing up the two values.
3. Extracted date from timestamp.
4. Aggregated data: aggregated data by multiple dimensions based on the analysis requirements, e.g. date, certain entities.

The processed data has 19,969,563 tweets in total.

4.1 Temporal Analysis

TweetsCov19 contains tweets data from Oct 2019 to Dec 2020. Tweets data is aggregated by date to generate the number of tweets on each date. As shown in the figure below, the number of tweets drastically increased in March 2020. The spike happened on 12th Mar, 2020 and the number of tweets on that day is 104,723.

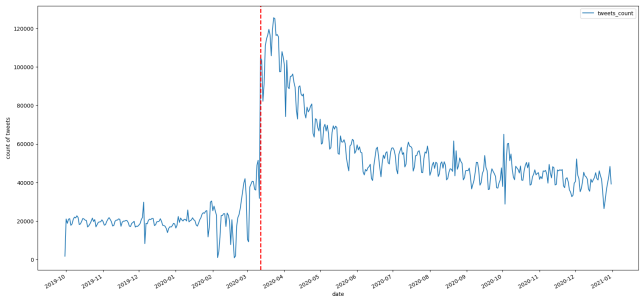


Figure 1: **Number of tweets over time**

On 11th Mar, 2020, World Health Organization(WHO) declared Covid 19 as a pandemic and Donald Trump, president of the United States, announced Covid 19 as a national emergency and unlocked billions of dollars of federal funding to fight the disease's spread. These two events could probably explain the cause of the sharp increase in the number of tweets. The frequency of hashtags *#coronavirus* and *#covid19* in tweets also displays the same pattern as shown in Figure 2. The usage of these two hashtags on 12th Mar 2020 increased by more than three times compared the the previous day.

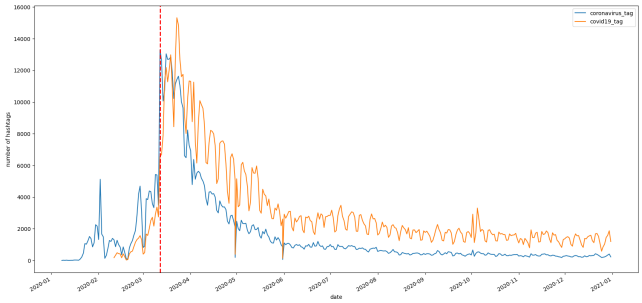


Figure 2: **Frequency of hashtag #coronavirus and #covid19**

The following plot shows the change of average sentiment score of tweets over time. The lowest average sentiment score is on 10th May 2020, which is -0.2211. A comparison of sentiment score before and after Mar 2020 has been done and summarized in Table 1. The overall sentiment score after Mar 2020 tends to be a lot more negative compared to the one before Mar 2020. The 75th percentile of average sentiment score is still negative for tweets after Mar 2020 while for tweets before Mar 2020, the 50th percentile is already above 0.

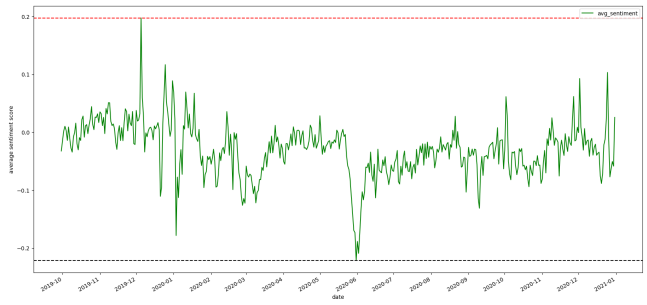


Figure 3: **Average sentiment score over time**

Period	Mean	25th %	50th %	75th %
Before Mar 2020	-0.0070	-0.0302	0.00213	0.0172
After Mar 2020	-0.0407	-0.0604	-0.0342	-0.0178

Table 1: **Average sentiment score before and after Mar 2020**

Besides sentiment changes over time, public opinions are monitored by measuring relationships between retweets and sentiment and exploring frequent itemsets in tweets. Based on the given positive and negative sentiment score in original data, summation of them is calculated to get the overall sentiment score ranging from -4 to 4. a score of -4 means the tweet content is extremely negative, a score of 4 means extremely positive, and a score of 0 implies neutral sentiment.

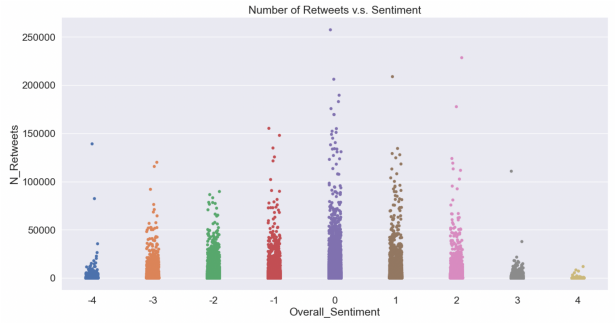


Figure 4: **Relationship between retweets and sentiment**

The Figure 4 above shows the relationship between number of retweets and overall sentiment scores. Each dot presents a single tweet, and all tweets are grouped by their sentiment scores. Tweets with neutral sentiment scores are dominant and the one that has the highest number of retweets belongs to this group. Another finding lies in groups with extreme sentiment scores. Compared to positive tweets, there are more negative tweets about COVID-19 and they reach a higher number of retweets, which matches the temporal analysis implying increasing of negativity after announcement of COVID-19.

Popular entities are explored in extreme tweets (tweets with sentiment scores of -4 and 4). The Figure 5 below presents two word clouds that contain top 100 entities in extreme positive or negative tweets. Bigger size of the word means higher frequency. Besides common words such as 'covid19' and 'quarantine', positive tweets mention more about love or romance, and negative tweets mention more about racist or political opinions. It can be assumed that people's negative emotions are related to government policies during COVID-19 such as lockdown and vaccination.

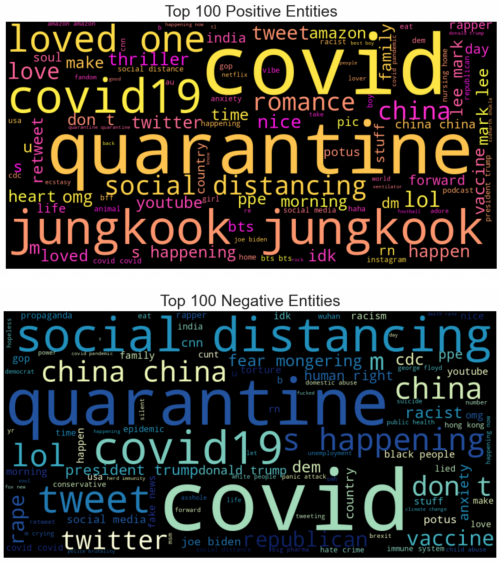


Figure 5: **Top 100 entities mentioned in extreme positive or extreme negative tweets**

In addition to the above analysis, an Auto-Regressive Integrated Moving Average (ARIMA) model is conducted to better explore the trend of change in average sentiment scores. The autocorrelation plot (Figure 6) shows that positive correlation exists with the first 10 lags and indicates that the first 5 lags are possibly significant. Models of order from 3 to 10 have been built and results show that when order is 3, the p-values of all lag variables are significant. The normality of residuals is also satisfied. The model summary can be found in Figure 7 and its predicted value plot and residuals distribution can be found in **A ARIMA MODEL RESIDUAL PLOT**.

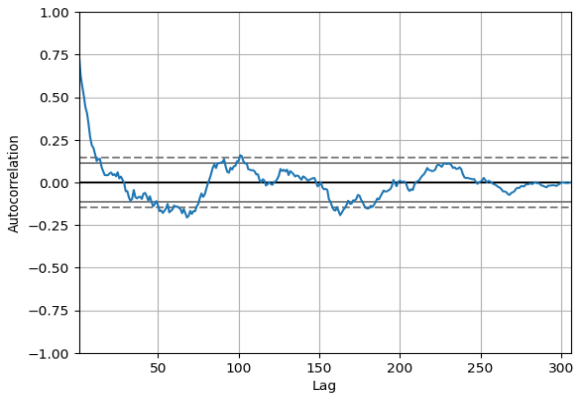
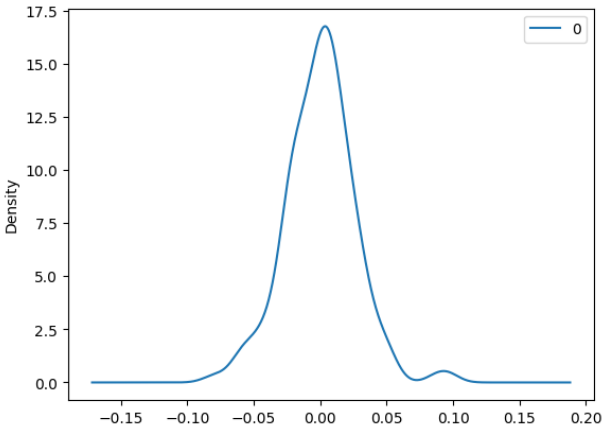


Figure 6: **Autocorrelation plot of sentiment score**

```
==============================================================================
Dep. Variable:          avg_sentiment   No. Observations:                  306
Model:                 ARIMA(3, 1, 0)   Log Likelihood                678.959
Date:                Sat, 29 Oct 2022   AIC                          -1349.918
Time:                        12:54:55   BIC                          -1335.037
Sample:                             0   HQIC                         -1343.966
                                  - 306
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3005      0.042     -7.148      0.000      -0.383      -0.218
ar.L2         -0.2298      0.047     -4.921      0.000      -0.321      -0.138
ar.L3         -0.1069      0.054     -1.982      0.047      -0.213      -0.001
sigma2         0.0007   4.27e-05     15.984      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                31.61
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.19
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.53
===================================================================================
```

Figure 7: **ARIMA model summary**



A ARIMA MODEL RESIDUAL PLOT

4.2 Influencer Analysis

The first thing we could find about the influencer is – influencers tend to tweet much more than those original people. On average, each person sends about 2 tweets, but from Figure 8 we can see that users with more followers tend to tweet more and the increase on the average tweets is quite smooth and steady, though the speed seems to decrease as the users' follower increases. Users with more than 10 thousand followers send 20 tweets on average and as the number of followers increases, when it comes to users with more than 500 thousand followers, the average tweets they send becomes 49.

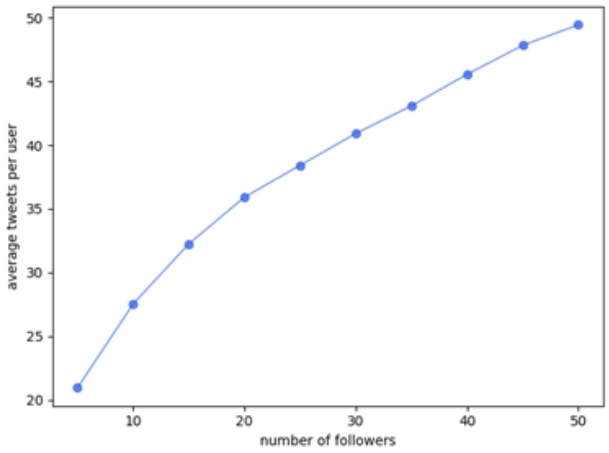


Figure 8: **Relation plot between average tweets and follower numbers**

Secondly, we can explore whether influencers are more neutral when making statements. From the bar chart of the sentiment of different kinds of people(Figure 9), we can find that there are no significant differences between intermediary people with followers between 100 and 100 thousand and ordinary people with followers less than 100. But for influencers with followers larger than 500 thousand, we can find that those influencers tend to be less word sentiment oriented. They tweet without personal emotions and their analysis and comments can be more fair than the other two groups of people. Meanwhile, influencers say much less extreme negative words than other people. Perhaps it is because that kind of words may influence the object being commented and they will be punished or even be indicted if they wrongly convey that kind of information.

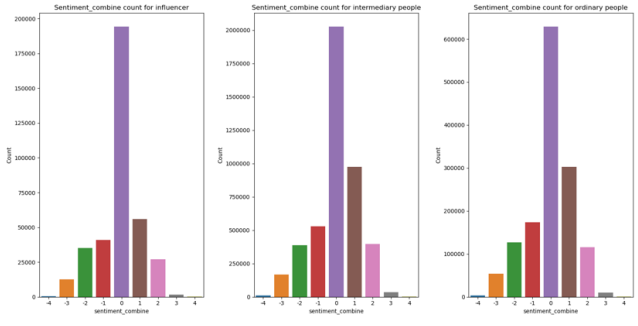


Figure 9: **Sentiment plot for different groups of people**

To see whether the influencers are using more hashtags and mentions in tweets, We use statistical methods to verify the differences between different groups. Since the distributions of these variables are not normally distributed and asymmetry but independent, we use Kruskal-Wallis test to compare between the difference of hashtag and mentions between different groups. From the result we can see that ordinary people, intermediary people and influencers are significantly different in using hashtags and mentions. Influencers tend to use less hashtags(0.57) than intermediary people(0.79) while intermediary people use less hashtags than ordinary people(0.84). It may be because influencers know more about how to use the most appropriate tags for their tweets so they will not abuse using them. Besides, too many tags in tweets will make readers annoyed and decrease the follower numbers and the influence of the influencer, therefore, influencers will try their best to choose the most precise tags for their tweets.

The same thing happens to mentions too. Influencers tend to mention less people(0.28) than intermediary people(1.23) on average and intermediary people mention less people than ordinary people(1.38). For influencers with hundreds of thousands of followers, their tweets are less likely to be used for personal affairs, the only possibility they mention someone was when they are taking advantage of celebrity interactions or mention their sponsors or advertisers. In this way, mentions will not be used very often by those influencers, especially large influencers. However, ordinary people can mention their friends and families whenever they want so they will mention more than influencers on average.

To see whether influencers use more normative statements, we construct entity null ratio as a statistical indicator. For the entity null ratio of different groups of people, we can find that influencers' tweets have much smaller entity null ratio (0.25) than the intermediary people (0.31) and the ordinary people (0.32). Given that the entities are extracted using a fast entity linker – an entity extraction tool using Wikipedia as a knowledge base, the entity null ratio indicates the normativity of the text. The more normative the tweets are, it will be easier for the fast entity linker to find entities in the tweets and therefore, the entity null ratio will become smaller. So we can see from our analysis that influencers' tweets are more normative than ordinary people.

4.3 **Frequent Itemset Analysis**

Lastly, analysis of frequent itemsets is applied on the data. Due to size of data and restriction of hardware capability, random sampling and Apriori method is taken to find the frequent itemsets. During the first step, ⅓ of the whole data has been utilized to filter out the frequent itemsets. Then, in order to prevent the issue of false positive, the occurrences of these shortlisted frequent itemsets are examined in the whole dataset to confirm that they are truly frequent.

Since random sampling is used instead of the whole dataset, there is a possibility that there exists false negative. i.e. There are frequent itemsets that are not picked during the first step. In order to minimize this issue, the support threshold has been set very low. In practice it is set as 0.02% of the whole data. Following is the result of the most frequent itemsets:

Itemset	Support
{Covid-19}	1304616
{China}	570715
{China, Covid-19}	13240
{Covid-19, Vaccine}	13327

Table 2: **Frequent itemsets and their support**

After we obtain the result of frequent itemsets, we examine whether there is any relationship between entities within the itemset. Hence, we conducted a temporal analysis in terms of popularity and sentiment of the entities. We use the itemset {China, Covid} as an example for illustration.

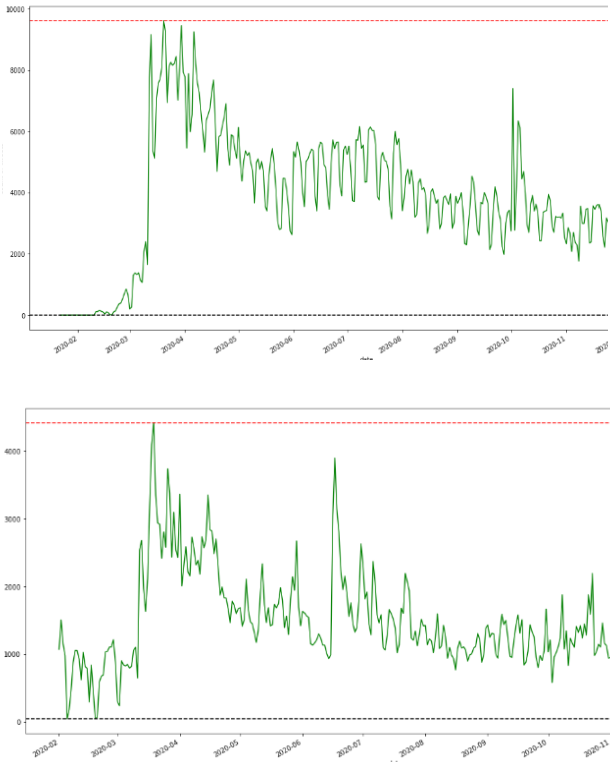


Figure 10: **Number of tweets of "Covid-19" and "China"**

The graphs reflect that the trend of number of tweets across time is similar for the 2 entities “Covid-19” and “China”. Both have sharp increases in occurrence in March 2020. Then both fall back, but still remain relatively high in popularity in tweets. During June to July 2022, there is another smaller peak for both entities.

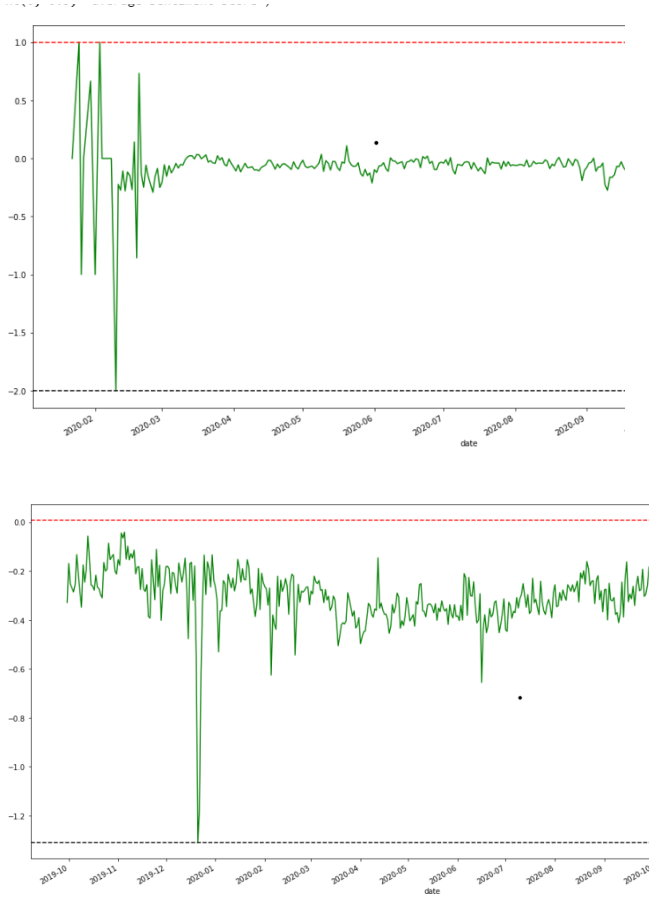


Figure 11 :**Negative Sentiments of “Covid-19” and “China”**

Above is the figure of negative sentiments of the entities across time. Both “Covid-19” and “China” have a sudden rise in extent of negative sentiment around the start of 2020, and remain slightly negative afterwards

From the figures, we can see that there is a very similar trend between entities within the same frequent itemset, in terms of both number of tweets and sentiment. Although we may not confirm there is strong causation relationship between the entities, it is still evident that there exists some correlation within frequent itemsets. Reversely, we may also make a hypothesis that entities with similar trends are likely to be a frequent itemset. More efforts are needed to delve into relations within frequent itemsets.

5 **Conclusion**

From our analysis, we find that firstly, the number of tweets of covid increased drastically after March 2020, and the overall sentiment is more negative as compared to before that time. That may be because there are not many Twitter users in China and not until March 2020, most of the users in Twitter started to realize the severity of the Covid, especially after the WHO and Donald Trump’s announcement. So during this period of time, the related tweets increased a lot and the sentiments of people are pessimistic.Secondly, using the ARIMA model we fit the sentiment trend and we find the sentiment in Twitter don’t last that long and it indicates the short-memory of the Internet. We also noticed that negativity increases the possibility of retweeting, and entities within the same frequent itemset may share a similar pattern in number of tweets and sentiment. That kind of pattern reflects the emotional transmission mechanism on the Internet. Lastly, we found that influencers have more tweets, are more neutral, more normative, and use less hashtags and mentions.

REFERENCES

[1] Dimitar Dimitrov , Erdal Baran , Pavlos Fafalios , Ran Yu , Xiaofei Zhu , Matthäus Zloch , and Stefan Dietze. 2020. TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. *In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM ’20), October 19–23, 2020, Virtual Event, Ireland. ACM, New York, NY, USA, 8 pages.* https://doi.org/10.1145/3340531. 3412765

[2] Jiménez-Zafra SM, Sáez-Castillo AJ, Conde-Sánchez A, Martín-Valdivia MT.2021. How do sentiments affect virality onTwitter? R. Soc. Open Sci.8: 201756.https://doi.org/10.1098/rsos.201756

[3] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. J. Am. Soc. Inf. Sci. Technol. 61, 12 (December 2010), 2544–2558.

[4] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and Space-Efficient Entity Linking for Queries. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15). Association for Computing Machinery, New York, NY, USA, 179–188. https://doi.org/10.1145/2684822.2685317

[5] Aasish Pappu, Roi Blanco, Yashar Mehdad, Amanda Stent, and Kapil Thadani. 2017. Lightweight Multilingual Entity Extraction and Linking. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17). Association for Computing Machinery, New York, NY, USA, 365–374. https://doi.org/10.1145/3018661.3018724