Data Article

Dataset of discourses about COVID-19 and financial markets from Twitter

Vu Minh Ngo

School of Banking, Business College, University of Economics Ho Chi Minh City, 59C Nguyen Dinh Chieu st, District 3, Ho Chi Minh City, Vietnam

ARTICLE INFO

Article history:
Received 2 February 2022
Revised 16 June 2022
Accepted 24 June 2022
Available online 30 June 2022

Dataset link: COVID-19 and financial markets tweets dataset (Original data)

Keywords:
COVID-19 pandemic
Financial markets
Cryptocurrency markets
Tweets

ABSTRACT

In this data article, a collection of 11,625,887 tweets on the topic of the COVID-19 pandemic are provided. The data from Twitter were collected through Twitter API from January 2020 to June 2020. In addition, we also provided subsets of tweets containing discourses on both COVID-19 and financial topics. In order to facilitate the research on sentiment analysis, the Sentiment140 dataset containing 1,600,000 tweets that were annotated as positive or negative sentiment was also provided (Go et al., 2009) We used Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to transform documents to numeric vectors and used logistic regression classifier to train and predict sentiments of tweets. These datasets may garner interest from data science, economists, social science, natural language processing, epidemiology, and public health groups.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

E-mail address: vunm@ueh.edu.vn

https://doi.org/10.1016/j.dib.2022.108428

2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Specifications Table

| Subject                           | Social Science |
|----------------------------------|----------------|
| Specific subject area            | Text mining and sentiment analysis to explore and understand most discussed COVID-19 and financial market-related topics and their effects on people's life during the first few months of the pandemic. |
| Type of data                     | R-Data (\*.rds file) |
|                                  | R code (\*.R file) |
| How the data were acquired       | We used Twitter API search with Academic Research access level |
| Data format                      | Raw |
| Description of data collection   | We searched on COVID-19 related keywords such as "coronavirus", "COVID-19", "corona", "social distancing", and "quarantine" from January to June 2020. All tweet texts were cleaned to remove any non-text elements using self-made python code and were saved in R data type to be read for sentiment analysis in R. The subset of COVID-19-financial markets topics are filtered using keywords of traditional finance topics such as "forex", "finance", "financing", "bank", "banking", "stock", "trading", etc. or keywords of cryptocurrency topics such as "blockchain", "ethereum", "altcoin", "ripple", "xrp", "crypto", "insurtech", "crowdfunding", "e-wallets", etc. Data are provided together with sample R and Python code to illustrate how to process, filter tweets, and conduct sentiment analysis. |
| Data source location             | University of Economics Ho Chi Minh City |
|                                  | Address: 59C Nguyen Dinh Chieu st, District 3, Ho Chi Minh City, 70,000, Vietnam |
|                                  | Sentiment140 dataset (clean_tweet.csv). Source: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip |
| Data accessibility               | Mendeley Data |
|                                  | Data identification number: 10.17632/4nc28dk9f3 |
|                                  | Direct URL to data: https://data.mendeley.com/datasets/4nc28dk9f3 |
| Related research article         | V. M. Ngo, & H. H. Nguyen. Are fear and hope of the COVID-19 pandemic responsible for the V-shaped behavior of global financial markets? A text-mining approach, Applied Economics Letters,(2021), 1–11. 10.1080/13504851.2021.1904105 |

Value of the Data

- These data are valuable for identifying and monitoring most topics discussed related to the COVID-19 pandemic during their first 6-month of breaking out, which could be used to aid in the development of interventions and communication to achieve desired behavior changes in response to the COVID-19 subject on Twitter.
- From this general COVID-19 dataset of tweets, we provided dataset samples and methods to filter tweets that specifically discussed financial market topics using suggested keywords. Thus, the dataset could also be helpful in studying people's reactions to financial issues in the context of COVID-19.
- These data will be beneficial for scholars doing both qualitative and quantitative studies on how the pandemic is perceived via social media postings. Scholars could use natural language processing such as text mining techniques to identify frequent themes and topics related to the COVID-19 pandemic.
- In addition, we also provided a robust and straightforward method of machine learning to extract sentiment data from the dataset of tweet texts. Moreover, different machine learning and natural language processing techniques could be applied to confirm and add more insights into the sentiment analysis. The sentiment analyses of tweets provide researchers and practitioners valuable inputs to study the impacts of the COVID-19 pandemic on many aspects of people's lives.
- Many Twitter datasets about the COVID-19 are available such as [3–5]. However, most of them focus on understanding the public sentiment or interactions between users on social networks when discussing only the COVID-19 pandemic in general. Our dataset is very few
that focus on the intersections between the COVID-19 pandemic and financial markets. It allows potential users to understand the impact of the COVID-19 pandemic on different financial markets, including conventional financial sectors and cryptocurrency sectors.

1. Data Description

There was a total of 11,625,887 tweets collected which was in the raw dataset (tweet_COVID_raw.rds) for each month from January to June 2020. Then the dataset of tweets collected was cleaned to keep only text (remove URL, emotion symbols, etc.) and filtered to remove duplicated tweets in terms of text, and saved as R-type datasets (*.rds file) for tweets in all languages (5,647,758 tweets) and tweets in only English (3,421,709 tweets).

We used another set of keywords and methods specified in the file “Filter finance-crypto tweets.R” to create two subsets of English tweets on traditional finance topics (42,315 tweets) and cryptocurrency topics (5563 tweets) in the context of the first six months since the COVID-19 pandemic broke out. These two datasets of financial markets-related tweets also have the same format as all other tweet datasets mention above.

We also provided some code files in R as examples and illustrations of how to use the tweet dataset to conduct the sentiment analysis. The “TFIDF_train.R” code file is used together with “clean_tweet.csv” which contain sentiment annotated tweets (0 = negative, 4 = positive) to train models. The “TFIDF_sentiment_prediction.R” code file is used for predicting the probability of positive sentiment of a tweet based on the trained model. Cleaned tweet datasets could be used with this code file to produce sentiment prediction for each tweet in the dataset. Tweet datasets, code files, and other files are described in Table 1.

| Dataset name          | Description                                                                 | Variables                                                                 |
|-----------------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------|
| tweet_COVID_raw.rds   | Collected tweets using related COVID-19 keywords                             | created_at: full date and time of tweets created                           |
|                       | Approx size: 934 MB                                                         | full_text: original text of the tweets                                     |
|                       |                                                                             | retweet_count: number of retweets received by the tweet                   |
|                       |                                                                             | lang: languages of the tweets                                             |
|                       |                                                                             | mention_name: Twitter account name mentioned in the tweet                 |
|                       |                                                                             | date: the date when tweets created                                        |
|                       | Tweet dataset whose text was cleaned and filtered to remove duplicated      | ID: tweets ID                                                             |
| tweet_COVID_clean.rds | tweets for sentiment analysis and natural language processing analysis.     | retweet_count: number of retweets received by the tweet                   |
|                       | Approx size: 249 MB                                                        | lang: languages of the tweets                                             |
|                       |                                                                             | date: the date when tweets created                                        |
| tweet_COVID_clean     | Clean tweet dataset with only English                                       | ID: tweets ID                                                             |
| en.rds                | in the tweet text.                                                         | retweet_count: number of retweets received by the tweet                   |
|                       | Approx size: 147 MB                                                        | lang: languages of the tweets                                             |
|                       |                                                                             | date: the date when tweets created                                        |
|                       |                                                                             | clean_text: cleaned text of tweets                                        |

(continued on next page)
The distribution of the number of tweets (including retweets) over time in the dataset of "tweet_COVID_raw.rds" is presented in Fig. 1 together with the number of daily COVID-19 confirm case growth [9] from January to June 2020. It could be seen that the number of tweets varies together with the dynamic of daily COVID-19 new confirmed cases. The topic of the COVID-19 pandemic started to gain interest on Twitter at the end of January 2020 and after WHO declared it as a global pandemic on 11th March 2020 [10]. The discourse about COVID-19 slowed down during April and May when the new cases growth was stable. However, at the end of May 2020, when the new cases growth picked up again, discussion about the COVID-19 topic on Twitter also significantly increased.

Similar trends of discussion about financial markets on Twitter are also observed in Fig. 2. The average number of daily tweets that mention both COVID-19 and traditional finance topics is 302 in the dataset. This number for tweets that mention both COVID-19 and cryptocurrency/fintech topics is much lower, with about 40 tweets a day.
Fig. 1. Number of COVID-19 related tweets (including retweets) overtime and COVID-19 daily case growth from January to June 2020. Source: [1].

Fig. 2. Number of traditional finance and cryptocurrency related tweets (including retweets) overtime from January to Jun 2020. Source: [1].
2. Experimental Design, Materials and Methods

2.1. Data Collection

Due to the COVID-19 virus' global spread, social media platforms have evolved into locations of intensive and constant information flow between government agencies, specialists, and the general public. Numerous scientific studies have shown that social media and news websites may be valuable sources of data for crisis analysis as well as for understanding people's attitudes and behavior during a pandemic [6,7].

To aid in public health surveillance and to aid experts in making decisions, numerous monitoring systems have been created to categorize enormous volumes of data from social media. This data may be used to rapidly detect the ideas, attitudes, sentiments, and subjects on which people's minds are focused in response to the COVID-19 epidemic [11]. Systematic study of this data may assist policymakers and health professionals in identifying problems of public concern and resolving them in the most effective manner.

We collected tweets about the COVID-19 discourse on Twitter from the end of January to early June 2020. To get tweets collected, we use a developer account in Twitter which allows us to access Twitter API v2 to search tweets on specific topics. With the access option for the academic research project, we were able to search the archive database of tweets on Twitter based on keywords and date specifications. We use code samples from Twitter Developer Relations and the package of search-tweets-python to aid us in tweet collections (https://github.com/twitterdev/search-tweets-python). All the tweets were collected using a set of COVID-19 related keywords, including "coronavirus", "COVID-19", "corona", "social distancing", and "quarantine". These keywords were chosen based on their popularity on Google Trends when regarding to the topic of COVID-19 at the time of data collection. The parameters for the tweet search include data fields such as 'author id', 'username', 'created_at', 'geo', 'id', 'lang', 'like_count', 'quote_count', 'reply_count', 'retweet_count', 'source', 'tweet'. The search result will return JSON format data containing all these predefined data fields. Then, we transformed JSON files into R data type files (rds files) for further processing and analysis using R. In the end, because of the availability and the purpose of this data is for sentiment and topic modeling, the final tweet datasets include 7 data fields as described in Table 1 (tweet_COVID_raw).

2.2. Corpus' Content Analysis

The effects of the COVID-19 pandemic could be explored by understanding what terms people mentioned the most in their tweets. Table 2 shows the 25 English terms that occur the most in tweets collected in "tweet_COVID_clean_en", "tweet_COVID_finance", and "tweet_COVID_cryptocurrency" datasets. In this analysis, stop words library and common words related to COVID-19 pandemic such as "coronavirus", 'covid', "covid19", "people","virus", etc. are excluded to explore other related topics besides the pandemic themes.

In the "tweet_COVID_clean_en" dataset, the most used terms are mostly about the impacts of the COVID-19 pandemic on people's health and lives. Governments' policy to deal with the COVID-19 pandemic could be another important discourse in this dataset of tweets. In the "tweet_COVID_finance" dataset, impacts of the COVID-19 pandemic on people's financial situation could be one of the main discourses based on the most used terms. Lastly, in the "tweet_COVID_cryptocurrency" dataset, some important topics in cryptocurrency and fintech are popular cryptocurrencies such as Bitcoin (btc), Ethereum (eth); new trends and technologies in fintech such as crowdfunding, insurtech or blockchain, etc.

In addition to terms' frequency, relationships and networks between terms could also be explored using the cleaned texts in tweet datasets. Fig. 3 shows one example of a network map of 100 most used bigram terms in the "tweet_COVID_clean" dataset. As presented in Fig. 3, "health" is the most connected term as the COVID-19 pandemic raised a lot of concerns about health
Table 2
Top 25 terms used in the three tweet datasets.

| terms          | tweet_COVID_clean_en | % of total tweets | terms          | tweet_COVID_finance | % of total tweets | terms          | tweet_COVID_crypto | % of total tweets |
|----------------|----------------------|--------------------|----------------|---------------------|--------------------|----------------|-------------------|--------------------|
| china          | 292,123              | 7.7%               | bank           | 7787                | 12.7%              | bitcoin        | 2209              | 17.8%              |
| lockdown       | 182,828              | 5.1%               | china          | 6085                | 3.9%               | crypto         | 1511              | 2.2%               |
| trump          | 147,431              | 3.9%               | stocks         | 5531                | 3.1%               | blockchain     | 1217              | 1.3%               |
| health         | 107,548              | 3.0%               | insurance      | 4862                | 4.8%               | china          | 1071              | 1.1%               |
| time           | 106,948              | 3.0%               | cash           | 4813                | 3.5%               | fintech        | 637               | 2.8%               |
| stay           | 92,516               | 2.4%               | debt           | 3771                | 1.5%               | cryptocurrency | 631               | 1.1%               |
| news           | 82,906               | 2.3%               | gold           | 3330                | 2.8%               | btc            | 539               | 5.8%               |
| outbreak       | 80,737               | 2.3%               | trading        | 2533                | 2.9%               | news           | 342               | 4.1%               |
| social         | 79,658               | 2.3%               | loan           | 2438                | 1.3%               | ripple         | 339               | 0.7%               |
| day            | 79,131               | 2.1%               | investment     | 2404                | 1.9%               | market         | 281               | 1.3%               |
| government     | 72,731               | 2.0%               | health         | 2268                | 2.9%               | xrp            | 280               | 1.3%               |
| wuhan          | 71,976               | 2.0%               | finance        | 2265                | 0.8%               | money          | 249               | 1.5%               |
| deaths         | 61,323               | 1.6%               | money          | 2145                | 0.6%               | eth            | 238               | 0.2%               |
| distancing     | 60,316               | 1.7%               | lockdown       | 1702                | 0.6%               | outbreak       | 233               | 0.3%               |
| chinese        | 60,177               | 1.6%               | time           | 1595                | 2.0%               | digital        | 218               | 2.7%               |
| spread         | 56,977               | 1.6%               | market         | 1562                | 1.5%               | markets        | 198               | 1.5%               |
| country        | 56,928               | 1.6%               | business       | 1508                | 1.5%               | binance        | 193               | 0.5%               |
| death          | 52,322               | 1.4%               | trump          | 1492                | 0.6%               | global         | 191               | 0.5%               |
| president      | 50,238               | 1.4%               | invest         | 1483                | 1.0%               | gold           | 173               | 0.3%               |
| public         | 48,745               | 1.4%               | economy        | 1457                | 1.4%               | business       | 170               | 0.6%               |
| days           | 46,255               | 1.3%               | global         | 1367                | 1.0%               | price          | 169               | 0.8%               |
| care           | 46,046               | 1.3%               | government     | 1255                | 1.4%               | technology     | 165               | 0.5%               |
| india          | 44,136               | 1.1%               | news           | 1250                | 1.6%               | time           | 160               | 0.9%               |
| cdc            | 43,242               | 1.2%               | investing      | 1230                | 0.6%               | trading        | 159               | 0.3%               |
| crisis         | 42,288               | 1.2%               | pay            | 1158                | 2.6%               | ethereum       | 157               | 2.3%               |

Fig. 3. Network map of most popular bi-terms used in the “tweet_COVID_clean” dataset. Source: [1].
issues such as public health, mental health, health crisis, or the need for health insurance or health care service.

2.3. Corpus’ Sentiment Analysis

As mentioned above, we included some R code files in the dataset as examples and demonstrations on how to do sentiment analysis using a Twitter dataset. To train models, the "TFIDF train.R" code file and the "TFIDF sentiment prediction.R" code file are used to estimate the likelihood of a tweet being positive in sentiment. Fig. 4 presented an example of sentiment analysis for a sample of 10,000 random tweets from the "tweet_COVID_finance" dataset. Each point in Fig. 4 presents the positiveness score for each tweet using the sentiment analysis of the TFIDF algorithm. Even though it is not obvious, we could observe that the number of tweets that have a probability of positive sentiment over 0.5 is larger than the number of tweets that have a higher probability of negative sentiment.

Conducting the TFIDF sentiment analysis for all three tweet datasets, the distribution of positiveness score for tweets in the three datasets is presented in Fig. 5. Sentiment distributions of the conventional finance and cryptocurrency datasets show that it is more likely that the positiveness score of a tweet is more than 0.5 indicating the positive sentiment of the tweet. In contrast, negative sentiment is more likely for a tweet in the COVID-19 dataset according to the distribution of sentiment score.
Fig. 5. Distribution of sentiment in the three datasets.

Ethics Statements

All data were collected and distributed under Twitter's Developer Policy 2021 [8].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

COVID-19 and financial markets tweets dataset (Original data) (Mendeley Data).
CRediT Author Statement

Vu Minh Ngo: Conceptualization, Methodology, Software, Visualization, Writing – review & editing.

Acknowledgment

This work was supported by the University of Economics Ho Chi Minh City (UEH) [2022-01-29-0817]

References

[1] V.M. Ngo, H.H. Nguyen, Are fear and hope of the COVID-19 pandemic responsible for the V-shaped behavior of global financial markets? A text-mining approach, Appl. Econ. Lett. (2021) 1–11, doi: 10.1080/13504851.2021.1904105.

[2] A. Go, Bhayani, R. and Huang, L. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford, (2009) 1(12). Accessed at https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[3] F. Haouari, M. Hasanain, R. Suwaileh, T. Elsayed, Arcov-19: the first arabic COVID-19 twitter dataset with propagation networks. arXiv preprint (2020) arXiv:2004.05861, doi: 10.48550/arXiv.2004.05861.

[4] H. Dashtian, D. Murthy, CML-COVID: a large-scale COVID-19 twitter dataset with latent topics, sentiment and location information. arXiv preprint (2021) arXiv:2101.12202, doi: 10.48550/arXiv.2101.12202.

[5] R.K. Gupta, A. Vishwanath, Y. Yang, COVID-19 Twitter dataset with latent topics, sentiments and emotions attributes. arXiv preprint (2020) arXiv:2007.06954, doi: 10.48550/arXiv.2007.06954.

[6] CSSECOVID-19 Data Repository By the Center for Systems Science and Engineering (CSSE) At Johns Hopkins University, CSSE, 2021. https://github.com/CSSEGISandData/COVID-19.

[7] WHOWHO Director-General’s Opening Remarks at the Media Briefing on COVID-19, WHO, 2020 https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[8] Twitter, Developer agreement and policy, (2021). https://developer.twitter.com/en/developer-terms/agreement-and-policy.

[9] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, S. Sharif, An analysis of COVID-19 vaccine sentiments and opinions on Twitter, Int. J. Infect. Dis. (2021) 108, doi: 10.1016/j.ijid.2021.05.059.

[10] K.M. Ridhwan, C.A. Hargreaves, Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore, Int. J. Inf. Manag. Data Insights 2 (2021) 100021, doi: 10.1016/j.ijimi.2021.100021.

[11] S. Liu, &J. Liu, Public attitudes toward COVID-19 vaccines on english-language Twitter: a sentiment analysis, Vaccine 39 (2021) 5499–5505, doi:10.1016/j.vaccine.2021.08.058.