COVID-19 RUMORS

Mingxi Cheng[1], Songli Wang[2], Xiaofeng Yan[3,†], Tianqi Yang[4,†], Wenshuo Wang[5,†], Zehao Huang[1], Shahin Nazaian[1], and Paul Bogdan[1,*]

[1]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, CA, USA, 90007
[2]School of Computer Science, Fudan University, Shanghai, China, 200433
[3]Department of Automation, Tsinghua University, Beijing, China, 100084
[4]University of Michigan - Shanghai Jiao Tong University Joint Institute, Shanghai, China, 200240
[5]Department of Automation, Beihang University, Beijing, China, 100083
[†]These authors contributed equally to this work
[*]Corresponding author: P.B. (pbogdan@usc.edu)

ABSTRACT

The coronavirus disease (COVID-19) has been rapidly spreading all over the world since December 2019. Along with the COVID-19 pandemic and the increasing reliance on online information sources, we witnessed an increasing rumor dissemination and misinformation activity. Rumors represent the diffusion of information accompanied by various interpretations and comments. Many research groups have released large scale datasets based on rumor content and spread trend analysis. However, utilizing the data in advanced research efforts involving higher-level language processing, such as sentiment analysis and intent detection, may require additional data integration. Labeled and clean data can serve research groups on various topics. For these purposes, we build a COVID-19 rumors dataset. We present 6834 rumors collected from Twitter and several news websites via automatic crawlers. The collection process runs from March to May 2020. All the rumors are tracked with source, post date, and popularity. We manually label each rumor with sentiment, veracity, and stance. Moreover, rumors in our dataset have 36230 retweets and 32750 reposts; hence, we provide stance labels to these records as well. Our dataset is free for research use and can serve well for projects relating to rumor detection, natural language processing-based rumor analysis, source tracking of misinformation, and measurement of public sentiment towards the pandemic.

Background & Summary

The outbreak of the coronavirus disease (COVID-19) has caused widespread concern among the public and has profoundly impacted social opinions since December 2019[1]. The ongoing pandemic has aroused heated discussion on the Internet[2]. All sorts of information are disseminated in the form of news reports, tweets, etc[3]. Topics of public discussion vary over time as the virus spreads rapidly, while real and fake information is mixed leading to increasing confusion in some communities. Rumors are defined as statements or reports currently without known veracity concerning their truthfulness,[4] which spread misinformation/disinformation (deliberately misleading information) and cause panic, hatred, and discrimination[5]. Along with the rapid dissemination of misinformation, researchers discover a massive growth in fact-checks about COVID-19, e.g., the number of English-language fact-checks increased more than 900% from January to March 2020[6]. Fact-checking websites such as the FactCheck.org[7] and Poynter.org[8], are the primary source of current COVID-19 misinformation/rumor data.

COVID-19 rumor datasets have been collected, e.g., the COVID-19-TweetIDs dataset contains an ongoing collection of tweets IDs associated with the COVID-19[9]; the COVID-19-Arabic-Tweets-Dataset contains a collection of Arabic tweets IDs related to the COVID-19[10]; the CoAID (Covid-19 healthcare misinformation dataset) is a diverse COVID-19 healthcare misinformation dataset, including fake news on websites and social platforms, along with users' social engagement with such news[11]. Besides the veracity labels and sources provided in the above-mentioned fact-checking platforms, other meta-information, such as sentiment and stance, is missing in most datasets. Therefore, COVID-19 rumor datasets for the study of sentiment analysis and other rumor classification tasks, including stance verification of COVID-19 rumors, are still in great need. In this work, we collect COVID-19 rumors and manually label 6834 data (4129 rumors from news and 2705 rumors from tweets) with sentiment and stance labels. The veracity status of the rumors is collected from the fact-checking websites. Moreover, we include 32750 reposts for news rumors and 34847 retweets for Twitter rumors and manually labeled the stance. Examples of our data structure and statistics are shown in Figure 1 and Table 1. Besides, we provide simple analysis of our dataset, including the statistical analysis of the rumor dissemination phenomena and some results on a deep learning-based rumor classification analysis.

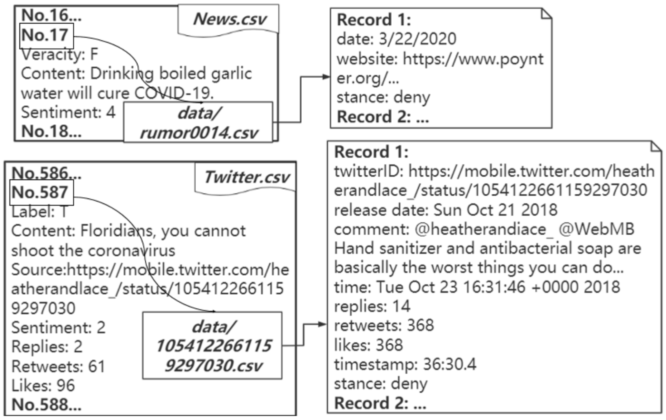

Figure 1. Examples of data structure.

Dataset	Records	Month	Replies	Frequency
News	4129	Jan.	3164	9.6%
		Feb.	5413	16.5%
		Mar.	15310	46.7%
		Apr.	1895	5.8%
		Others	6968	21.4%
Tweets	2705	Jan.	12702	35.1%
		Feb.	9350	25.8/%
		Mar.	12646	34.9%
		Apr.	1386	3.8%
		Others	146	0.4%

Table 1. Summary of data records.

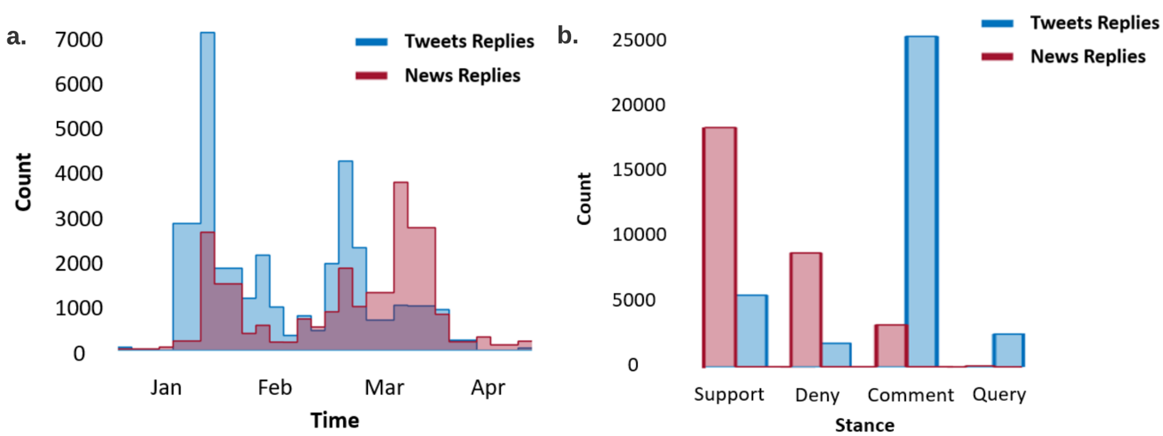

Figure 2. Reply records and their stance statistics in the collected COVID-19 dataset.

Methods

We collect rumors from a wide range of sources and refer to various authoritative fact-checking platforms (for their comprehensive analysis), e.g., *poynter.org*, *factcheck.org*. We also collect online discussions from Twitter, which contained real-time discussions with specific tags. To gather information from Twitter, we focus on several relevant tags and official accounts (e.g., ABC News, Reuters, CNN, and BBC News) to trace the updates of the hot topics. We record the rumor sentences in our dataset and enrich the dataset with more details, such as the source website, date of publication, veracity, sentiment, and stance. We also include reposts or retweets of the rumors and provide their stance labels. We separate our collected rumors into two datasets based on their source: (*i*) a News dataset, containing rumors collected from news websites, and (*ii*) a Twitter dataset, containing rumors collected from Twitter.

Data Collection

We develop web crawlers to extract data automatically from the Google browser and Twitter. The codes are all available on our GitHub repository.

Tweets Collecting Method

We collect tweets with COVID-19 related tags, such as *COVID-19*, *coronavirus*, *COVID*, and store them in *.csv* files in the format of *ID_url*, *release_date*, and *full_text*. Duplicated tweets are dropped. Then, the sentiment of each rumor is labeled through careful analysis of the emotion of the rumor content and context. Moreover, we fetch metadata of each tweet, including source websites, reply/retweet comment content, reply numbers, retweet numbers, likes, and publish dates. These data are saved in individual files named by the tweets' IDs. Then we manually label the stance of the replies or retweet comments.

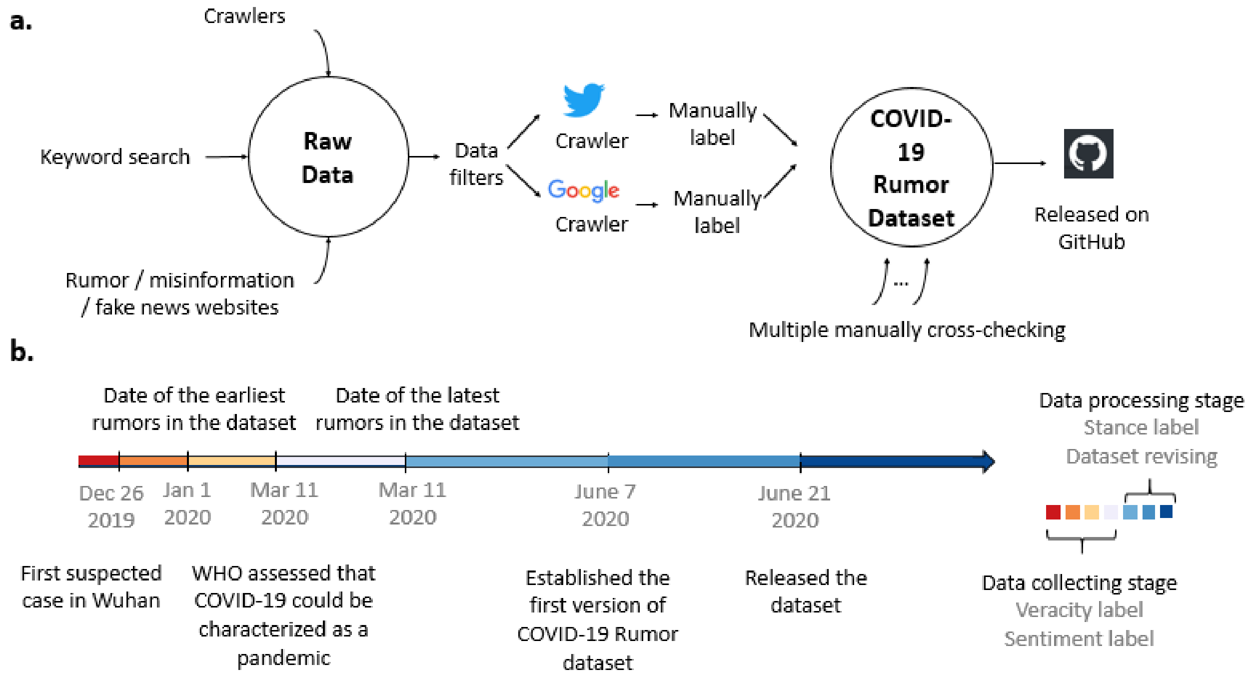

Figure 3. a. Dataset collection, labeling, and post-process flowchart. **b.** Timeline of dataset establishment and major events in the outbreak of the COVID-19 pandemic.

News Collecting Method

We use the *mitmproxy*[12], an open source interactive HTTPS proxy, to collect detailed information related to selected news reports from Google browser. Unlike the Twitter crawler, the Google crawler mainly catches data from the search result page and queries the results' absolute and relative paths according to the *URL* and date. We save the news rumors in a *news.csv* file, and the repost records are saved in *rumorID.csv*'s. Each rumor record has a veracity label and the rumor content, and each repost record has a repost date, repost website, and stance label. It is worth noting that not all source website contains date information.

Data Records

The full data is freely available on GitHub. All records are stored in *.csv* files. Utf-8 encoding is recommended for the best display. We build two datasets to store the rumors from news and Twitter, respectively. Metadata are provided for further research use an analysis; meanings of the labels are demonstrated in Table 2.

News Dataset

The news dataset contains rumors from news reports about COVID-19, including emergency events, comments of public figures, updates on the coronavirus outbreak, etc. Each record contains the following formatted metadata describing the details of the news:

Sources. Websites containing the rumor sentence. Note that web pages discussing the rumor, e.g., discussing the veracity of the rumor, are counted as sources, and the earliest rumor source is noted as the origin.

Popularity. The popularity of each news is the number of all websites in Google browser that repost the whole rumor.

Date. The published date of each rumor record, which is collected by the web crawler automatically.

Stance. The attitude of the author or editor of the rumor source. We follow classical rumor stance classification and define four classes of stance: support, deny, comment, and query[13]. The stances are labeled and cross-validated manually by going through the context of each website. It's worth noting that the stances are concentrated in support and comment types. Stance statistics of both news and Twitter set are shown in Figure 2.

Sentiment. We define a rumor sentence with five fine-grained types of sentiment: very negative, negative, neutral, positive, and very positive. We manually label and cross-validate the sentiment according to whether it's good news or bad news. For

Term	Label	Demonstration
Veracity	True (abbreviation: "T")	The content is logical and describing the facts, e.g."Wuhan has been quarantined."
	False (abbreviation: "F")	The content is made up, or contains false information, e.g., "Drinking bleach can cure coronavirus."
	Unverified (abbreviation: "U")	The authenticity or truthfulness of the statement is hard to judge at the time of labeling.
Stance	Support	Positive attitudes about the content, e.g."I think the statement is right."
	Deny	Denying attitudes about the content, e.g."Are you kidding? This is wrong!"
	Comment	No obvious stance, e.g."This message is interesting."
	Query	Doubting the validity of news/tweets, e.g."Is that true?" or "Can you prove?"
Sentiment	Very Negative	The content has a strong pessimism.
	Negative	The emotion is pessimistic but weaker than "very negative"
	Neutral	The comment/report is in a plain and narrative tone.
	Positive	The content reflects positive emotions or aims, such as news providing tips to fight the virus.
	Very Positive	Cheerful news such as progress in the research, massive donations or breakthroughs in the vaccine.

Table 2. Demonstration of labels of each rumor in the dataset.

example, news reporting new infections is usually labeled as negative; news indicating death due to COVID-19 would be labeled as very negative; news providing tips of virus prevention is more likely to be labeled as positive; news reporting progress in the scientific research, massive donations or breakthroughs in the vaccine would be rated as very positive.

Veracity. The veracity status of each rumor can be true, representing the news is describing a fact; or false, representing the news is a false claim; or unverified, indicating the news cannot be verified by the time of the collection. The labelling is conducted and cross-validated manually at the data collecting stage based on authoritative websites and shared common knowledge.

Twitter Dataset

The Twitter dataset contains rumors posted on Twitter. The data are grabbed from public accounts, which comment on the updates of COVID-19 related information discussion under certain tags. The discussions may then be retweeted or quoted by other accounts.

Source. For each tweet, we obtain the source (origin) through twitter ID and content. The format of a tweet's source is https://twitter.com/user/status/ID.

Reply. Replies indicate how people respond to the tweet, which are similar to sources in News dataset.

Reply/Retweet/Like (RRL) number. These numbers show the propagation trend of a tweet. The RRL number is enclosed in Twitter JSON and parsed automatically by the crawler.

Popularity. Popularity of a tweet is represented by the sum of RRL numbers.

Date. The date is when the tweet is published on the Twitter, which is represented by the format: MM.DD.YYYY.

Stance. Stance represents the attitude of the sources quoting the original tweet or people commenting on it. We provide stance for rumor itself and its retweets.

Sentiment. We label and cross-validate the sentiment according to the emotion of the tweet. In addition, online tools of sentiment-analysis like MonkeyLearn[14] are used for comparisons. We find that compared to our manual assessment, the out-of-box tools make a significant amount of mistakes and lead to low accuracy.

Veracity. The definition is the same as in news dataset.

Technical Validation

We manually label and cross-validate all records by referring to multiple sources. Veracity is imported and cross-checked based on multiple fact-checking websites, and we take the most agreed labels. For the records collected from public sources such as Snopes, Politifact, and Boomlive, we conduct a second manual validation step to verify its reliability. For the remaining data crawled from news websites and social media, we manually review each and every record by multiple people. For each record, if two labels are inconsistent, an extra labeling step will be done by a 3^{rd} person to ensure the labeling quality.

For sentiment labeling, we manually label and verify by five people using the same standards. Additionally, two out-of-box tools, flair[15] and MonkeyLearn[14], are used to cross-validate the outcomes. Flair is a natural-language-processing(NLP) python package from `http://github.com/flairNLP/flair`, and MonkeyLearn is an online sentiment analysis tool based on natural language processing. Given an input sentence, flair outputs a positive/negative sentiment with a score ranging from 0 to 1. We further evenly divide these ranges into afore-mentioned 5 categories, and compare the results with our manual labels, and adjust them accordingly. On the other hand, for each input sentence, MonkeyLearn sentiment classifier picks one tag from the collection (very positive, positive, neutral, negative, very negative), which can be projected directly to the five categories defined in Data Records Section. We find that machine-generated labels are not accurate compared to manual labels. Here are two typical examples:

Figure 4. Sentiment comparison examples.

Usage Notes

Our dataset contains COVID-19 rumors and its metadata, such as veracity, sentiment, popularity, etc. In this section, we present a potential data analysis and rumor classification tasks that could be done using our dataset as follows:

- **Rumor characterization.** We analyze the statistical laws that characterize the rumor popularity data and find that the log-normal distribution cannot be denied as the data generating process.
- **Rumor classification.** We provide a case study of rumor classification tasks with our collected dataset including veracity classification, stance classification, and sentiment analysis.

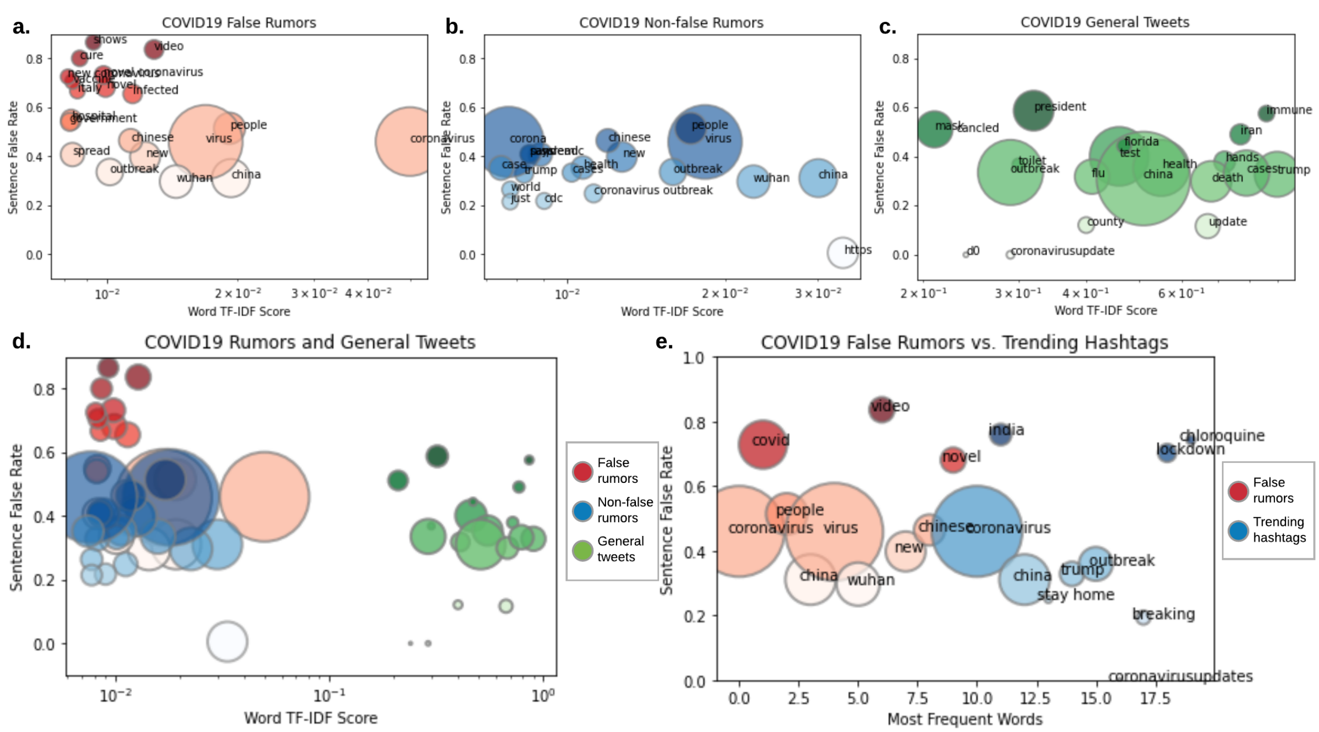

Figure 5. TF-IDF vs. sentence false rate. **a-d.** Comparison between top words in COVID-19 tweets, false rumors, and non-false rumors. **e.** Comparison between frequent words in false rumors and trending Twitter hashtags. Color and size of the bubble indicate word source and frequency.

Top words and trending hashtags in general tweets have different veracity patterns.

We compare the *top words* in our rumor dataset and general COVID-19 tweets. Words with the highest term frequency-inverse document frequency (TF-IDF) scores[16] are referred to as *top words*. We find that words in Twitter posts with the highest TF-IDF scores share very similar *sentence false rate* as top words in our non-false rumors (non-false represents true and unverified rumors). *Sentence false rate* of a word in this work is calculated as the ratio of the occurrence of the word in false rumor sentences and occurrence of the word in any rumors (both false and non-false). If the word 'wash' appears in false rumors f times and appears in non-false rumors n times, then the sentence false rate of 'wash' is calculated as $f/(f+n)$. A high sentence false rate indicates a word appears much more frequently in false rumor than in non-false rumors. Figure 5a-d illustrate the comparison between the TF-IDF and sentence false rate of top words in false rumors, non-false rumors, and general COVID-19 tweets. We find that top words in tweets appear almost equally in false rumors and non-false rumors, i.e., the green bubbles in Figure 5c have around 0.5 sentence false rate. Top words in our non-false rumors have the same patterns as top words in general COVID-19 tweets, as shown in Figure 5b, i.e., the blue bubbles have around 0.5 sentence false rate. The sentence false rate similarity between top words in general COVID-19 tweets and our non-false rumors indicates that the general veracity/truthfulness of COVID-19 Twitter posts are towards non-false.

Different from top words in general Twitter posts and non-false rumors, we find that top words in our false rumors (marked as red bubbles) have different patterns, as shown in Figure 5a. The red bubbles lying in a much higher sentence false rate range, $[0.4, 1.0]$, indicate that the top words in false rumors, such as 'cure' and 'vaccine', appear much more frequently in false rumors than in non-false rumors. These words are 'false rumor-specific', meaning that other words that are frequently seen in non-false rumors, such as 'outbreak' and 'cases', do not share the same significance as 'cure' or 'vaccine' in false rumors.

After comparing the top words in general COVID-19 tweets and our rumor dataset, we analyze the trending coronavirus related Twitter hashtags and find that the steady trending hashtags behave similarly to *the most frequent words* in false rumors. *The most frequent words* are those that appear most frequently in some content, in our case, false rumors. As shown in Figure 5e, the trending hashtags in Twitter (marked as blue bubbles), such as 'india', 'lockdown', 'chloroquine', and most frequent words in false rumors (marked as red bubbles), such as 'covid', 'video', 'novel', lie in similar sentence false rate range, meaning that the veracity of trending hashtags is at the same level as false rumors. Compared to top Twitter words, trending Twitter hashtags have higher sentence false rates, which leads to the warning that Twitter posts with trending hashtags have a higher risk of being false rumors than general COVID-19 Twitter posts. In summary, top words in general tweets share the same pattern as top words in non-false rumors while trending Twitter hashtags share the same pattern as the most frequent words in false rumors.

Statistic laws governing rumor popularity.

Researchers have noticed for decades that many measured data retrieved from biological and social sciences can be described by log-normal distribution[17,18] and power-law distribution[19]. In this case study, we estimate the log-normal and power-law models of our collected rumor popularity data, including both news dataset and Twitter dataset. We utilize *powerlaw* package in python to estimate the fitted models[20]. We perform a statistical hypothesis test analysis as follows: (i) We estimate the parameters of fitted models via *powerlaw*. For example, for the fitted power law model, the estimated parameter α indicates the index of power law's probability distribution: $p(x) \propto x^{\alpha}$; and the estimated x_{min} indicates the optimal start point of the power law fit (tail). (ii) After parameter estimation, we calculate the goodness-of-fit between the popularity data and the power-law (and log-normal). Specifically, we calculate p_{KS}, a plausibility value, based on measurement of the "distance" between the distribution of the empirical data and the hypothesized model. The distance D is estimated by *powerlaw* when we fit the data, which is the "Kolmogorov-Smirnov (KS)" statistic. Next, we generate a large number of power-law (and log-normal) synthetic data with the estimated parameters and we fit the synthetic data using *powerlaw*. After fitting the synthetic data, we get the distance of synthetic data and the hypothesized power-law model (and log-normal model fitted by the synthetic data), noted as D_{syn}. Then we repeat this procedure by generating 100 sets of synthetic data with 100 D_{syn}'s. Finally we calculate p_{KS} as the percentage of $D_{syn} > D$. If p_{KS} is greater than 0.1, the power-law (or log-normal) is a plausible hypothesis for the data[21]. (iii) After calculating the p_{KS}, we compare hypotheses, power-law and log-normal, via a likelihood ratio test provided in *powerlaw*, e.g., $R, p = distribution\_compare('lognormal', 'powerlaw')$, where R is the log-likelihood ratio between the two candidate distributions. If $R > 0$, then the data are more likely to follow the first distribution, otherwise the data are more likely to obey the second distribution. p is the significance value for that direction. The favored distribution is a strong fit if $p > 0.05$. Following the hypothesis test, our estimated results of rumor popularity data are shown in Figure 6.

The distribution fitting comparison between power law and log-normal for News and twitter data are shown in Figure 6. The estimation and the resulting parameters for the power law and the log-normal are shown in Figure 6a-b. The likelihood ratio test results for news and Twitter data are $R, p = (-3.0001, 0.0027)$ and $R, p = (-3.1880, 0.0014)$, respectively. The negative R values indicate that the data are more likely to follow the second distribution, i.e., log-normal. The goodness-of-fit tests are therefore performed for log-normal hypothesis and the results are shown in 6c-d. With p_{KS} values greater than 0.1 in all experiments, we conclude that the rumor popularity data (from both news and Twitter dataset) are indistinguishable from identically and independently drawn samples from the log-normal distribution.

Deep learning-based rumor classification tasks

We implemented a BERT[22](Bidirectional Encoder Representations from Transformers) and VAE[23](Variational Auto-Encoder) to build our deep-learning based rumor classification system. Our classification system is applied to three tasks: rumor veracity

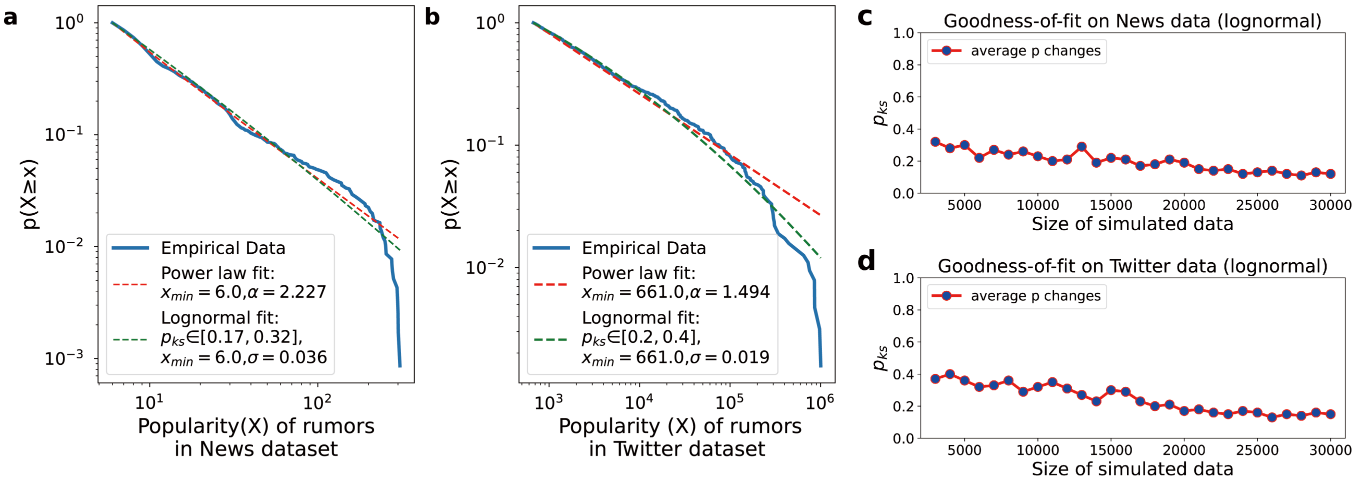

Figure 6. Comparison between log-normal and power-law fit of twitter rumors and the goodness of fit test on log-normal distribution. The p_{KS} on log-normal fit of Twitter data falls in $[0.20, 0.40]$. The result indicates the original data matches log-normal distribution well. The p_{KS} of news data falls in $[0.17, 0.32]$, which also means the original data matches log-normal distribution well.

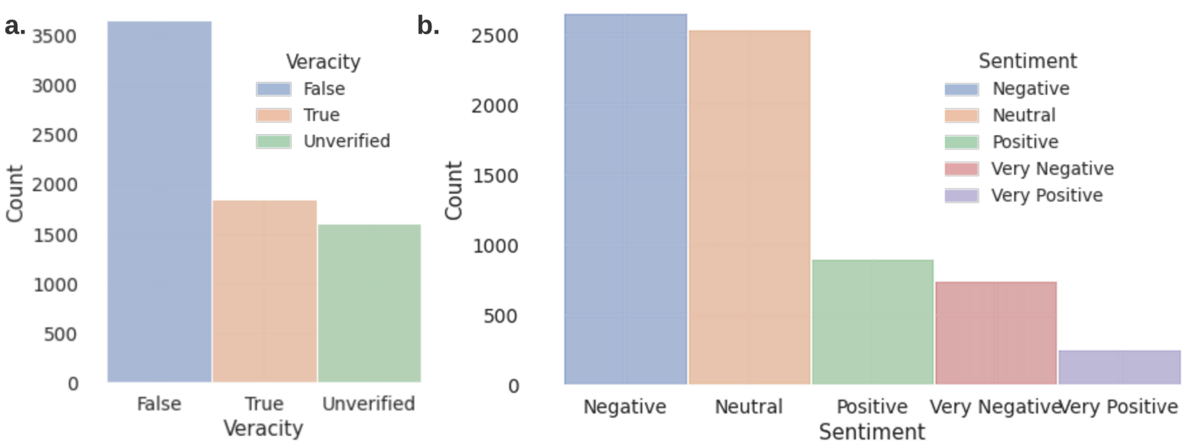

Figure 7. Veracity and sentiment statistics of our collected dataset.

classification, stance classification, and sentiment classification. The statistics of the veracity, sentiment, and stance in our dataset are shown in Figure 7 and Figure 2. In our experiment, BERT is fine-tuned to serve as a word embedding layer that converts rumor text to vectors. The BERT is a representation model pre-trained by jointly conditioning on both the left and right context in all layers. Therefore, deep bidirectional representations of unlabeled text on both left and right contexts in all layers can be learned. In the next step, we use an LSTM-based variational auto-encoder in VRoC[13] to extract features from the vectors generated by BERT. Finally, we input the feature vectors to a 3-layer fully connected neural network to carry out classification. In summary, the classification results (the F1-scores) of three tasks achieved by our model is 0.85, 0.77, and 0.47.

Code availability

All the code used for processing and the dataset are available on GitHub (`https://github.com/MickeysClubhouse/COVID-19-rumor-dataset`). The code is freely available and the dataset is in compliance with the Twitter's Terms and Conditions. The dataset is licensed under MIT License. Please refer to the README file in the code released for further information.

References

1. Nelson, L. M. *et al.* Us public concerns about the covid-19 pandemic from results of a survey given via social media. *JAMA internal medicine* (2020).
2. Cinelli, M. *et al.* The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004* (2020).
3. Kouzy, R. *et al.* Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus* **12** (2020).
4. Friggeri, A., Adamic, L., Eckles, D. & Cheng, J. Rumor cascades. In *Eighth International AAAI Conference on Weblogs and Social Media* (2014).
5. Depoux, A. *et al.* The pandemic of social media panic travels faster than the covid-19 outbreak (2020).
6. Brennen, J. S., Simon, F. M., Howard, P. N. & Nielsen, R. K. Types, sources, and claims of covid-19 misinformation. *Reuters Inst.* (2020).
7. Brooks, K. H. J., Eugene. Factcheck.org: A project of the annenberg public policy center (2020).
8. Neil Brown, K. M. The coronavirusfacts/datoscoronavirus alliance database (2020).
9. Chen, E., Lerman, K. & Ferrara, E. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372* (2020).
10. Alqurashi, S., Alhindi, A. & Alanazi, E. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315* (2020).
11. Cui, L. & Lee, D. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).
12. Cortesi, Maximilianhils & Raumfresser. Mitmproxy, an open source interactive https proxy. https://www.mitmproxy.org/ (2020).
13. Cheng, M., Nazarian, S. & Bogdan, P. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of The Web Conference 2020*, 2892–2898 (2020).
14. MonkeyLearn, I. Monkeylearn-text analysis. https://monkeylearn.com (2020).
15. Akbik, A., Blythe, D. & Vollgraf, R. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, 1638–1649 (2018).
16. Rajaraman, A. & Ullman, J. D. *Mining of massive datasets* (Cambridge University Press, 2011).
17. Sun, K. Explanation of log-normal distributions and power-law distributions in biology and social science. *Tech. Report, Dep. Phys.* (2004).
18. Limpert, E., Stahel, W. A. & Abbt, M. Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. *BioScience* **51**, 341–352 (2001).
19. Clauset, A. Trends and fluctuations in the severity of interstate wars. *Sci. advances* **4**, eaao3580 (2018).
20. Alstott, J., Bullmore, E. & Plenz, D. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one* **9**, e85777 (2014).
21. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).
22. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
23. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Author contributions statement

M.C., S.N., and P.B. contributed to the design of the research including data collection, data processing, data analysis, experiments, and organizing the information in the main text. M.C., S.W., T.Y., X.Y., Z.H., S.N., and P.B. contributed to the writing of the manuscript. S.W., T.Y., X.Y., and W.W. contributed to the data collection and labeling. W.W. contributed to the crawler implementation. M.C., S.W., and Z.H. contributed to the implementation of data analysis experiments. M.C. and X.Y. contributed to preparing the figures and tables, and their captions. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.