Detecting Chinese Fake News on Twitter during the COVID-19 Pandemic

Yongjun Zhang
Stony Brook University
yongjun.zhang@stonybrook.edu

Sijia Liu
Amazon.com, Inc
sijial@amazon.com

Yi Wang
Stony Brook University
yiw.wang@stonybrook.edu

Xinguang Fan
Peking University
xfan19@pku.edu.cn

Abstract

The outbreak of COVID-19 has led to a global surge of Sinophobia partly because of the spread of misinformation, disinformation, and fake news on China. In this paper, we report on the creation of a novel classifier that detects whether Chinese language social media posts from Twitter are related to fake news about China. The classifier achieves an F1 score of 0.64 and an accuracy rate of 93%. We provide the final model and a new training dataset with 18,425 tweets for researchers to study fake news in the Chinese language during the COVID-19 pandemic. We also introduce a new dataset generated by our classifier that tracks the dynamics of fake news in the Chinese language during the early pandemic.

1 Introduction

Alongside the COVID-19 pandemic, the spread of misinformation, disinformation, and fake news via social media platforms such as Twitter, Facebook, and Tiktok has led to the infodemic concerned by policymakers, journalists, academics, and the public (Buchanan, 2020; Germani and Biller-Andorno, 2021; Simon and Camargo, 2021; Zarocostas, 2020). Scholars have developed a series of computational tools to detect and monitor COVID-19 fake news on social media platforms, but prior studies focus overweeningly on the dominant languages in Western societies such as English and German (Bang et al., 2021; Mattern et al., 2021; Vijjali et al., 2020).

We focus on fake news detection in the Chinese language context on a western social media platform in this paper, given that China has become the epicenter of this infodemic due to the conspiracy on the origin of COVID-19 from a Wuhan lab. During the early pandemic, the misinformation and fake news on COVID-19 was first spread within the Chinese community and then diffused to other communities. However, we lack critical data to quantify and understand the scale and scope of the diffusion of fake news related to China during the early COVID-19 pandemic in its own community.

In this paper, we adopt a subjective opinion-based approach to detect fake news on Twitter in 2019-2020 (Zhang and Zadorozhny, 2020). Unlike the fact-check approach relying on third-parties (e.g., news media, government organizations, experts), we mimic the real world scenario about how fake news recipients perceive these messages and focus on social media users' subjective opinion of whether or not these messages are fake news.

Our contribution is threefold. First, we provide a unique annotated training dataset with over 18,000 tweets. To our best knowledge, this is the first training dataset in the community specifically related to fake news on China during the pandemic. Second, we fine-tune a series of transformer models and train our own classifiers that can be used to classify whether a Chinese tweet is fake news or not. Third, we also provide a dataset with over 25 million Chinese tweets obtained via Twitter academic API using keywords mentioning China, Chinese, CCP, etc. We provide an overview of the dynamics

of fake news on Twitter in the Chinese language since the outbreak of COVID-19.

2 Related Work

In recent years, both industrial and academic research scientists have devoted great attention to fake news detection due to the growing concerns regarding the adversarial consequences of fake news. As a result, multiple large-scale annotated fake news datasets have been created for algorithm training, but they are primarily in the English language (Mattern et al., 2021; Shu et al., 2017, 2019b, 2020). This allows scholars to train state-of-the-art deep learning models to detect and monitor fake news, and these methods account for both texual and contentual information (e.g., social media data) into fake news detection (Shu et al., 2019a; Zhou et al., 2020a; Zhou and Zafarani, 2019). Next, we provide a brief overview of prior fake news research related to the concept of fake news, approaches of fake news detection, and datasets of fake news detection.

2.1 Defining Fake News

An accurate and operational definition of fake news is a big challenge for fake news research. Existing studies adopt various concepts and terms, such as clickbait (Chen et al., 2015), misinformation (Lazer et al., 2018), disinformation (Mattern et al., 2021), and false information (Shu et al., 2017). Egelhofer and Lecheler (2019) offer a two-dimensional framework to define fake news: fake news genre (describing the deliberate creation of pseudojournalistic disinformation) and fake news label (the instrumentalization of the term to delegitimize news media). Zhang et al. (2020) survey the fake news literature and provide two different definitions. For the broad definition, fake news is simply false information, while the narrow one restricts fake news within the intentionally false news published by news outlets. The underlying consensus behind current fake news detection techniques takes the broad approach that fakes news is at least false information.

2.2 Approaches to Studying Fake News

Methods of fake news detection can be summarized into three main approaches. *Fact-checking* approach relies on the text of questionable claims and trains models on labeled claims from online fact-checking resources, such as *classify.news* and *FactCheck.org* (Graves, 2018; Pathak and Srihari,

2019; Popat et al., 2018; Rashkin et al., 2017). This approach does not use explicit features and often fails to offer interpretable explanations. *Style-classification* approach focuses on the textual (context texts, images, videos, etc.) and contextual features (user profiles, user posts and responses, etc.) of news and uses machine learning or deep learning based models to detect fake news (Przybyla, 2020; Zhang et al., 2020). However, two issues arise from this approach: dependence on context and news contents. Dependence on context indicates that fake news style varies by time, languages, and regions, while heavy dependence on news contents may inspire countermeasures by fake news makers. *Network-based* approach relies on news cascade and spreading features of fake news (Liu and Wu, 2020; Zhang et al., 2018; Zhou and Zafarani, 2019; Zhang and Zadorozhny, 2020; Shu et al., 2019a). This approach uses direct or indirect information on news propagation to train traditional machine learning or deep learning models and, therefore, mainly applies to the task of detecting fake news which has been spread over a period of time. Other scholars often categorize different approaches into knowledge-based and propagation-based fake news detection as well. The former focuses more on determining the truthfulness of news time while the latter emphasizes the credibility of news sources.

2.3 Fake News Benchmark Datasets

A few benchmark general fake news datasets have been released publicly in recent years. For instance, several large-scale fake news related datasets have been used by scholars to build fake news detection algorithms, including *BuzzFeed-News*, *LIAR*, *BS Detector*, *CREDBANK*, *BuzzFace*, *FacebookHoax,FakeNewsNet* (see Shu et al., 2019b for a detailed review). These datasets focus on different dimensions of fake news, such as linguistic, visual , social , and spatiotemporal features.

To combat the COVID-19 infodemic, scholars have developed a series of COVID-19 specific fake news datasets, including *Covid19 FN* (Patwa et al., 2021), *FakeCoovid* (Shahi and Nandini, 2020), *Re-COVery* (Zhou et al., 2020b), CoAID (Cui and Lee, 2020), and *FANG-COVID* (Mattern et al., 2021). Of course, these datasets are overwhelmingly in English (Hangloo and Arora, 2021), except for *FANG-COVID* and *FakeCovid*. Given that content style varies across languages and societies, non-English data availability is crucial for the further

development and application of fake news detection techniques.

2.4 Fake News Detection

Scholars have developed a variety of machine-learning or deep learning models to detect fake news (see Patwa et al., 2021 for a more comprehensive review). Here we briefly describe some models trained on the above datasets we have mentioned. For instance, Glazkova et al. (2021) fine-tuned Covid-Twitter-BERT model (Müller et al., 2020) and achieved an F1 score of 0.987. Shahi and Nandini (2020) obtained an F1-score of 0.78 based on English articles from FakeCovid with a pre-trained BERT model. Zhou et al. (2020b) trained a neural network that combines textual and visual information and achieved an F1-score of 0.672 for the detection of fake news using ReCOVery fake news data. Mattern et al. (2021) trained a deep learning model combining BERT with social context information based on *FANG-Covid* dataset with 41,242 German news articles and associated twitter information and achieved an accuracy of 0.981 and an F1 score of 0.966. But other models trained on general fake news datasets are not quite effective, which warrants further research on specific topics.

3 Introducing the Datasets

Next, we introduce two datasets used in this paper: CNTweets data and CNFakeTweets data. The former contains all Chinese tweets mentioning China, Chinese, Chinese Communist Party, Chinese government, etc. The latter is the training dataset we compiled for fake news detection on Twitter.

3.1 CNTweets Data

To evaluate how fake news in the Chinese language was diffused across communities, we rely on the CNTweets, a dataset containing all Chinese tweets mentioning China-related keywords, such as China, Chinese, Chinese Communist Party, etc. This dataset contains over 25 million tweets by 1.32 million users scraped from the Twitter historical database using academic API. CNTweets data has 16.84 million tweets mentioning China and 8.39 million mentioning CCP.

3.2 CNFakeTweets Data

To build our training dataset *CNFakeTweets*, we started with those Twitter users who are most/least likely to post fake news related to China and their followers or following accounts, and these accounts typically either support or criticize the Chinese government or China. We obtained all their tweets posted in the past 2 years. We also used pro- and anti-China hashtags and keywords to obtain potential tweets. We then used a stratified sampling strategy to select 12,000 tweets. To add more potential neutral tweets in our training dataset, we also randomly selected 3,000 tweets from the CNTweets data. These 15,000 tweets were annotated by at least two research assistants.

We then use the annotated training dataset to train our algorithm to classify fake news in our CNTweets database. We take a semi-supervised machine learning approach to construct our final training dataset. After fist-stage of classification, we annotate another random sample of 5,000 tweets classified by our algorithm as fake news. We then feed these tweets into our training dataset for our final classifier building. Thus, our final training dataset contains 18,435 tweets after data cleaning with 2,599 labeled fake news.

4 Data Annotation

We adopt a subjective opinion-based approach to annotate fake news in this study. We broadly define fake news as false claims that mimic news information and written intentionally to mislead readers (Shu et al., 2017; Zhang and Zadorozhny, 2020). We ask research assistants to read the content of a tweet solely based on its text and then evaluate whether it is fake news related or not based on their subjective reading and personal experiences. All tweets are evaluated at least by two annotators and we treat it a tweet as fake news as long as one annotator labels it as fake news.

5 Model Training and Performance

5.1 Transfer learning and fine tuning

We take a transfer learning approach to build our models to identify fake news in the Chinese language on Twitter during the early pandemic. Since previous studies consistently show that transformer models have outperformed other language models, we fine-tune a series of BERT models trained on Chinese corpora to categorize a tweet into fake news or not. We compare the results based on Chinese-BERT-WWM-ext, Chinese-RoBerta-WWM-ext, and Chinese-MacBert-large.

5.2 Model performance

Table 1 reports our model metrics. Our pioneer work on detecting fake news on Twitter in the Chinese language offers scholars some baselines for comparison. Note that we only account for textual information in a tweet.

Model Name	Accuracy	F1 score
Chinese-bert-wwm-ext	0.926	0.656
Chinese-roberta-wwm-ext	0.932	0.663
Chinese-macbert-large	0.929	0.648

Table 1: Transformer Model Performance

6 The Overview of Chinese Fake News on Twitter

Next, we present the overview of fake news in the Chinese language on Twitter during the early pandemic from 2019 to 2021. We use our best model to classify CNTweets data. Among over 25 million tweets, we find 0.97 million tweets classified as fake news. Roughly 4% of tweets posted in Chinese are classified as fake news.

6.1 Time Series of Fake News in Twitter

Fig. 1 shows the weekly trend of fake news in our CNTweets database. Overall, fake news persisted during the early pandemic. The early two spikes around February and May in 2020 were due to the conspiracy theory on the origin of COVID, and fake news on the Wuhan Lab was shared and retweeted by many Twitter users. The false information peaked around late 2020 and early 2021. Roughly 13% of tweets in our CNTweets database were classified as fake news. This is because misinformation on U.S. presidential election was spread on Twitterverse, for instance, the Hunter Biden laptop controversy.

6.2 Content of Fake News on Twitter

To further explore the contents of these fake news on Twitter, we run structural topic modeling to extract the major themes (Roberts et al., 2019). Fig. 2 shows the explored 20 topics in our database. The most prevalent topic in the extracted fake news pertains to COVID-19 conspiracy theories (e.g., made in Wuhan Lab and Limeng Yan's interview by Fox News) during the early pandemic. In addition, fake news posted in Chinese associated with U.S. presidential election warrants future scholarly attention. False information related to both Republican and
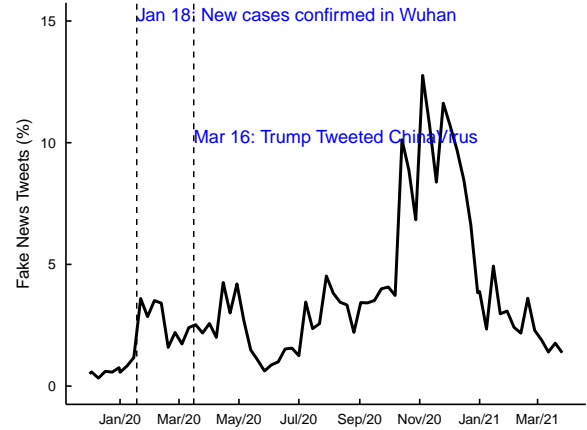

Figure 1: Weekly Trend of Tweets Labeled as Fake News in CNTweets Database
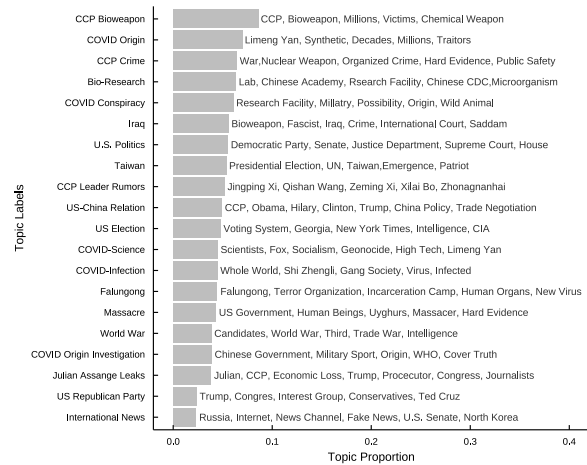

Figure 2: Structural Topic Model Results (K=20)

Democratic candidates in 2020 accounted for a large proportion in our fake news database.

7 Conclusion

Computer and social scientists have devoted great efforts to detect misinformation and fake news on social media platforms in recent years. Our paper provides the fist attempt, to our best knowledge, to systematically detect fake news in the Chinese language on a social media platform, i.e., Twitter. We annotated a unique training dataset with over 18,000 tweets used to build a transformer-based classifier. We then employed this classifier to a large-scale dataset during the early pandemic containing tweets mentioning any keywords related to China, Chinese, and Chinese Communist Party in the Chinese language. Our data visualization

shows a rise and fall of fake news on Twitter within the Chinese language community that needs further studies.

This paper adopts a subjective opinion-based approach to evaluate the credibility of a tweet, and our approach captures the subjective assessment of whether a tweet is a news-like false claim. Future research can further conduct fact-checking to verify the reliability of these subjective rating. In addition, this study does not incorporate Twitter-user or post-level auxiliary information, such as user profile, followers, following, and number of likes, into modeling. Future research can further improve our classifier by retraining BERT models with extra social media information.

Limitations

Our research has several limitations as well. First, our classifier only considered textual information in a tweet. As tweets posted by users are multimodal, researchers can take images, emojis, and videos into account in the future research. Second, previous research has shown that Twitter user information has substantially improved the accuracy of fake news classification. Although our work does not incorporate authors' information, we have collected users' timeline information and their networks and we will address this issue in the future work. Third, our classifier might not be extended to other periods since it focused on the early COVID pandemic. One approach to extend this line of work is to add more training examples in our CN-FakeTweets dataset. Lastly, we took an subjective approach to classify fake news, and one future direction would be to combine subjective approach with fact-check. It would be also important to compare the inconsistency between different annotators with intercultural backgrounds.

Ethics Statement

We used Twitter academic API to collect our CN-Tweets and CNFakeTweets datasets. Due to Twitter's term of service, we cannot share the original tweets we have collected, but we can share Tweet IDs and our classifiers. In addition to data ethics, our work also raises other ethical concerns. First, the training dataset labeled by research assistants have a lot of misinformation and racial slurs, which could cause some mental health issues. Scholars should prepare annotators with enough background and resources when they annotate the

training dataset. Second, our work only reveals the overall pattern in our collected database and these results do not reflect our personal opinions.

References

Yejin Bang, Etsuko Ishii, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2021. Model Generalization on COVID-19 Fake News Detection. *arXiv:2101.03841 [cs].* ArXiv: 2101.03841.

Mark Buchanan. 2020. *Managing the infodemic*. Ph.D. thesis, Nature Publishing Group.

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Jana Laura Egelhofer and Sophie Lecheler. 2019. Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association*, 43(2):97–116. Number: 2.

Federico Germani and Nikola Biller-Andorno. 2021. The anti-vaccination infodemic on social media: A behavioral analysis. *PloS one*, 16(3):e0247642.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at constraint@aaai2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 116–127. Springer.

Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. page 8.

Sakshini Hangloo and Bhavna Arora. 2021. Fake News Detection Tools and Methods – A Review. *arXiv:2112.11185 [cs].* ArXiv: 2112.11185.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Yang Liu and Yi-Fang Brook Wu. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33. Number: 3 Publisher: ACM New York, NY, USA.

Justus Mattern, Yu Qiao, Elma Kerz, Daniel Wiechmann, and Markus Strohmaier. 2021. FANG-COVID: A New Large-Scale Benchmark Dataset for

Fake News Detection in German. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 78–91, Dominican Republic. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Archita Pathak and Rohini Srihari. 2019. BREAKING! Presenting Fake News Corpus for Automated Fact Checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362, Florence, Italy. Association for Computational Linguistics.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29. Springer.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. *arXiv:1809.06416 [cs]*. ArXiv: 1809.06416.

Piotr Przybyla. 2020. Capturing the Style of Fake News. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497. Number: 01.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An r package for structural topic models. *Journal of Statistical Software*, 91:1–40.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Kai Shu, H. Russell Bernard, and Huan Liu. 2019a. Studying fake news via network analysis: detection and mitigation. In *Emerging research challenges and opportunities in computational social network analysis and mining*, pages 43–65. Springer.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019b. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *arXiv:1809.01286 [cs]*. ArXiv: 1809.01286.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36. Number: 1.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. In Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu, editors, *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, Lecture Notes in Social Networks, pages 1–19. Springer International Publishing, Cham.

Felix M Simon and Chico Q Camargo. 2021. Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic'. *new media & society*, page 14614448211031908.

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking. *arXiv:2011.13253 [cs]*. ArXiv: 2011.13253.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.

Danchen Zhang and Vladimir I. Zadorozhny. 2020. Fake News Detection Based on Subjective Opinions. In Jérôme Darmont, Boris Novikov, and Robert Wrembel, editors, *Advances in Databases and Information Systems*, volume 12245, pages 108–121. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B. Gouza. 2018. Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*.

Jiawei Zhang, Bowen Dong, and S. Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1826–1829. IEEE.

Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2020a. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25. Number: 2 Publisher: ACM New York, NY, USA.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020b. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 3205–3212.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60. Number: 2 Publisher: ACM New York, NY, USA.