A Study of Cantonese Covid-19 Fake News Detection on Social Media

Ziwei Wang*, Minzhu Zhao*, Yu Chen†, Yunya Song* Liang Lan‡
*Department of Journalism, Hong Kong Baptist University, Hong Kong SAR, China
†Department of Computer Science, Nanjing University, Nanjing, China
‡Department of Communication Studies, Hong Kong Baptist University, Hong Kong SAR, China
‡Corresponding author. E-mail address: lanliang@hkbu.edu.hk

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671722

***Abstract*—With the prevalence of social media, fake news has become one of the greatest challenges in journalism, which has weakened public trust in news outlets and authorities. During the COVID-19 epidemic, the widely circulated pandemic-related fake news on social media misleads or threatens the public. Recent works have investigated fake news detection on social platforms in English and Mandarin, though Cantonese fake news has been understudied. To pave the way for Cantonese COVID-19 fake news detection, we first presented an annotated COVID-19 related Cantonese fake news dataset collected from a popular local discussion forum in Hong Kong. Then, we explored the dataset by applying topic modeling to identify the topics that contain the most significant amount of fake news. Moreover, we evaluated both traditional machine learning algorithms and deep learning algorithms for Cantonese fake news detection. Our empirical results show that deep learning based methods perform slightly better than traditional machine learning methods on TF-IDF features.**

***Index Terms*—Fake News Detection, Topic Modeling, Cantonese Text Analytics**

I. Introduction

Fake news detection has gained increasing research attention in recent years [1]. However, Cantonese fake news has been understudied [2]. With this concern, we construct and annotate a Cantonese COVID-19 related dataset based on one of the most popular local discussion forums, LIHKG [1], in Hong Kong. Due to the immatureness of existing Natural language processing (NLP) technologies on Cantonese text analytics, it raises crucial questions about the applicability of existing fake news detection models on Cantonese. In this study, we investigate the following two research questions: **RQ1**: Are some topics containing more fake news than others? **RQ2**: How accurate are the existing fake news detection models on our Cantonese Covid-19 fake news dataset?

II. Methodology

We have framed this study with three main parts: data collecting and labeling, text data preprocessing, and fake news detection, as shown in Fig. 1.

[1]www.lihkg.com

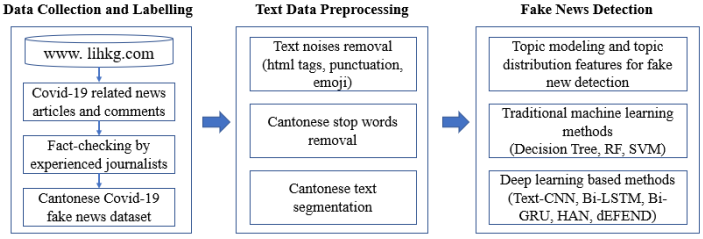

Fig. 1. The framework of our study on Covid-19 fake news detection

A. Dataset Collection and Labeling

Keywords based search were performed to collect the COVID-19 related posts from LIHKG firstly. The set of search keywords includes: "新冠" (abbr. New coronavirus), "肺炎" (Pneumonia), "Covid", "新型冠狀" (New coronavirus), "疫苗" (vaccine), "全民檢測" (Public testing), "口罩" (mask), "抗疫" (Anti-epidemic), "疫情" (Epidemic), "鍾南山" (Zhong Nanshan), "確診" (Confirmed positive), "張竹君" (Chuang Shuk-kwan), "袁國勇" (Yuen Kwok-yung). Then we developed a labeling system and recruited three experienced journalism student helpers to filter and label the data. Finally, we collected 1,917 posts along with comments for true and fake news. Table I provides an overview of the dataset.

B. Problem Setting for Fake News Detection

Let us use a_i to denote the i-th news article in the data and use $\mathbf{c}^i = \{c_1^i, c_2^i, \ldots, c_m^i\}$ to denote the m-comments associated with the news article a_i. And $y_i \in \{-1, 1\}$ is used to denote the label for a_i, for a news article a_i, $y_i = 1$ means a_i is fake and $y_i = -1$ indicates a_i is true. Then the fake news detection problem can be defined as follows. For a given dataset $\{(a_i, \mathbf{c}^i, y_i)\}_{i=1}^n$ containing n labeled news with their associated comments, we want to learn a fake news detection function such that $f(a, \mathbf{c}) = 1$ if a is fake and $f(a, \mathbf{c}) = -1$ otherwise.

To get the meaningful word representation for each news article, we applied three steps for text preprocessing: (1) noise removal of http tags, punctuation marks, and emojis; (2) stop words removal; and (3) Cantonese text segmentation.

TABLE I
OVERVIEW OF THE COLLECTED CANTONESE COVID-19 DATASET

	# of posts	# of comments	# of users
Fake-news	132	8,566	4,293
True-news	1,785	85,542	15,964
Total	1,917	94,108	20,257

C. Topics Exploration and Topic-related Features

To answer our **RQ1**, we propose to use the Latent Dirichlet Allocation (LDA) model [3] to mine the topics. Based on the LDA results, we will (1) identify the topics that contain significantly more fake news than other topics using the hypergeometric test; and (2) evaluate the performance of fakes news detection models by using the topic distribution as the feature representation. And here we only focus on the news articles $\{a_i\}_{i=1}^n$ and do not consider their associated comments.

By fitting the LDA model on our data $\{a_i\}_{i=1}^n$, we obtain a topic distribution for each news article, i.e., $p(a_i) = [\theta_1^i, \theta_2^i, \ldots, \theta_k^i]$ where k is the number of topics. The θ_j^i denotes the probability of news article a_i belonging to topic j ($j = 1, \ldots, k$). To identify the topics that contain significantly more fake news than other topics, we assign each news article to the topic with the highest probability θ_j^i. Based on the topic assignment and the label (i.e., true or fake) of the news articles, we use the hypergeometric test to determine whether some topics are over-represented by fake news.

In our experiments, we identified two topics containing significantly more fake news than others with p-values less than 0.01. The results suggest that the topic-related features could be meaningful indicators in fake news detection. Therefore, we also evaluate the fake news detection performance with topic distribution $[\theta_1^i, \theta_2^i, \ldots, \theta_k^i]$ as the feature representation for news articles.

D. Fake News Detection Algorithms

To answer our **RQ2**, we applied and evaluated both traditional machine learning based methods and deep learning based methods in our study.

1) Traditional machine learning based fake new detection: The general procedure for traditional machine learning based fake new detection contains two modules: feature extraction and classification model. For feature extraction, our study examined two content based features: (1) Bag-of-Word based Term Frequency-Inverse Document Frequency (TF-IDF) features: to explore the language difference; and (2) topic distribution based feature as described in II-C: to explore the topic distribution difference. The mentioned feature extraction methods can be consider as transforming the text data (a_i, c_i) into a numerical representation $\mathbf{x}_i$. After extracting features on news articles and comments, we transforming our raw data $\{(a_i, \mathbf{c}^i, y_i)\}_{i=1}^n$ to a standard input data format $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ for binary classification. For the classification model, we choose Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM) as the classifiers based on their demonstrated good performance in existing realted studies [1], [4].

2) Deep Learning based fake news detection: Unlike traditional machine learning algorithms that require a handcraft feature extraction process, the end-to-end deep learning models have been widely studied in text classification in recent years. Our study also applies and evaluates several popular deep learning based methods for text classification on our dataset. Specifically, we investigated text-CNN [5], Bidirectional Long Short-Term Memory Networks (Bi-LSTM) [6], Bidirectional Gated Recurrent Unit (Bi-GRU) [7], Hierarchical Attention Networks (HAN) [8] and dEFEND [9] in our experiments. The input word embedding of Cantonese to deep learning models are obtained by fastText [10].

III. EXPERIMENT

Topic Discovery. Table II shows the results of LDA with the number of topics equal to seven. The topic labels are assigned manually. Among seven topics, the topic of "vaccination" and "spoken Cantonese" contain a higher percentage of fake news than others with p-values less than 0.01.

Fake News Detection. Given our data is highly imbalanced, we use Area Under the ROC Curve (AUROC) and Area under the PR Curve (AUPRC) as the evaluation metrics. We randomly split the data into training set (80%) and test set (20%) to evaluate the performance. This procedure is repeated five times and the averaged AUROC and AUPRC are reported.

The results of different machine learning models are shown in Table III. Since our preliminary studies revealed that the comments data could not aid the performance while adding on extra time and source in our problem, Table III's results are based on news contents except dEFEND method with comments trails. Concerning traditional methods, features from topic distribution are more constructive than other features in our dataset for fake news detection. Under topic distribution features, RF gets the best AUPRC and SVM gets the best AUROC. For deep learning based methods, these methods generally get slightly better results than traditional machine learning methods with TF-IDF features. However, they get similar results as traditional machine learning with topics features. Deep learning based methods do not get significantly better results than traditional machine learning based methods in our dataset. The reasons could be (1) Our Cantonese COVID-19 data is relatively small; (2) The Cantonese text segmentation and Cantonese word embedding are immature as English and Mandarin.

IV. CONCLUSION

We introduced a labeled COVID-19 related Cantonese fake news dataset and explored three research questions

TABLE II
OVERVIEW OF THE TOPICS IN OUR DATASET

Topic id	Topic label	# of news	% of fake news	Top-10 representative words in the topic	Significant level
1	Quarantine	794	2.02	全民(mass), 樣本(sample), 計劃(plan), 員工(employee), 袁國勇(Yuen Kwok-yung), 疫(pandemic), 檢疫(quarantine), 逾(over), 健康(health), 港(Hong Kong)	1.000
2	Global pandemic	575	7.65	台灣(Taiwan), 年(year), 英國(United Kingdom), 報導(report), 澳洲(Australia), 特朗普(Donald Trump), 著(doing), 世界(world), 斯(Si: one popular word for names), (this)	0.220
3	**Vaccination**	**144**	**13.89**	**疫苗(vaccine), 抗體(antibody), 接種(vaccination), 劑(dosage), 注射(inject), 測試(test), 試驗(experiment), 團隊(team), 研發(Research and Development), 產生(produce)**	0.001***
4	**Spoken Cantonese**	**214**	**22.43**	**係(is), 唔(not), 嘅('s), 佢(he/ she), 你(you), 咁(this), 既('s), 大家(we), 話(say), 香港人(Hongkongers)**	0.000***
5	Police	63	1.59	警方(police), 警員(police officer), 影片(video), 警(police), 警察(policeman), 蕾(Lei: one popular word for names), 夏(summer), 槍(gun), 萊特(wright: names), 子女(children)	0.990
6	Medical Supplies	166	3.61	元(Taiwan), 生產(year), 億(100 million), 採購(buy), 銅(copper), 芯(core), 資助(fund), 企業(company), 供應(supply), 盒(box)	0.979
7	Health	10	10.00	戒(quit), 煙(smoking), 吸煙(smoke), 国(country), 嬲(tease), 蒸(steam), 思想(thought), 区(District), 发(supply), 比賽(competition)	0.511

TABLE III
COMPARISON OF DIFFERENT FAKE NEWS DETECTION MODELS

Model	Features	AUPRC	AUROC
DT	TF-IDF	0.12	0.68
RF	TF-IDF	0.16	0.68
SVM	TF-IDF	0.16	0.66
DT	topics	0.20	0.73
RF	topics	0.23	0.74
SVM	topics	0.21	0.77
text-CNN	content	0.14	0.74
Bi-LSTM	content	0.12	0.68
Bi-GRU	content	0.19	0.73
HAN	content	0.19	0.74
dEFEND	content + comment	0.21	0.73

based on this dataset. We found that two topics, "vaccination" and "spoken Cantonese", contain significantly more fake news than others. We evaluated both traditional machine learning and deep learning methods for fake news detection. We observed that topic distribution is a helpful indicator for classification. Our results also show that deep learning based methods perform slightly better than traditional machine learning methods on TF-IDF features in our dataset.

ACKNOWLEDGMENT

This work was supported by the Research Grant Council of Hong Kong (RGC/HKBU12605520), the Public Policy Research Funding Scheme (2021.A2.047.21B) from Policy Innovation and Co-ordination Office of the HKSARG, the Interdisciplinary Research Clusters Matching Scheme (IRCMS/19-20/D04) and the AIS Scheme (Ref. AIS 21-22/01) from Hong Kong Baptist University, and the Natural Science Foundation Council of China (61906161).

REFERENCES

[1] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[2] L. Ke, X. Chen, Z. Lu, H. Su, and H. Wang, "A novel approach for cantonese rumor detection based on deep neural network," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 1610–1615.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[6] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[7] M. Zulqarnain, R. Ghazali, M. G. Ghouse, and M. F. Mushtaq, "Efficient processing of gru based on word embedding for text classification," *JOIV: International Journal on Informatics Visualization*, vol. 3, no. 4, pp. 377–383, 2019.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[9] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.