An AI-based System to Assist Human Fact-Checkers for Labeling Cantonese Fake News on Social Media

Zi Hen Lin[‡], Ziwei Wang[†], Minzhu Zhao[†], Yunya Song[†] Liang Lan*[‡]
[†]Department of Journalism, Hong Kong Baptist University, Hong Kong SAR, China
[‡]Department of Interactive Media, Hong Kong Baptist University, Hong Kong SAR, China
*Corresponding author. E-mail address: lanliang@hkbu.edu.hk

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020949

***Abstract*—Preventing the spread of fake news is one of the most challenging issues in the age of social media. Traditional manual fact-checking (i.e., expert-based and crowd-sourced fact-checking) is time-consuming and labor-extensive, which cannot scale up with the unprecedented amount of dis- and mis- information on social media. Automated fact-checking based on machine learning is a promising strategy to address the scalability issues. Nevertheless, an end-to-end full automated fact-checking system without human supervision is still impractical. A more realistic solution will be developing an Artificial Intelligence (AI)-based system to facilitate the human fact-checkers during the fact-checking process. Therefore, this paper proposes a novel annotation system to facilitate human fact-checkers. With our designed procedures and schema, our developed system can help to improve the efficiency and effectiveness of human fact-checkers by automatically identifying worth-to-check news. We conduct a real-case study to demonstrate that our system can effectively identify worth-to-check news and ease the annotation process with the help of several automatic detection functions.**

***Index Terms*—Automatic fact-checking; Fake news detection; News annotation process and schema;**

I. Introduction

With the prevalence of using social media to create, distribute, share, and access news, people get exposed to a lot of fake news because that fake news can now be created and published faster and cheaper in social media compared to traditional news media [1]. For example, over one million tweets on Twitter related to the fake news story "pizzagate" were published by the end of the 2016 presidential election [1]. Ironically, recent research has shown that fake news on Twitter typically spreads far more rapidly than true news [2]. Fake news has become one of the biggest challenges in journalism, which has weakened public trust in governments and news outlets [3].

To combat fake news, manual fact-checking was initially developed in journalism. However, manual fact-checking is a time-consuming, intellectually demanding, and laborious process [4]. It cannot scale with the huge amount of newly created to-be-checked news information in social media. To address the scalability issue of manual fact-checking, machine learning-based automatic fact-checking methods [5] have been proposed in recent years. These studies have shown promising early results on automated fake news detection. These machine learning-based automatic fact-checking methods view fake news detection as a classification problem. That is, a to-be-checked news article can be represented as a feature vector $\mathbf{x}$ and the task is to predict the label y of this new article where y can be defined as $y = 1$ if it is fake and $y = -1$ if it is true.

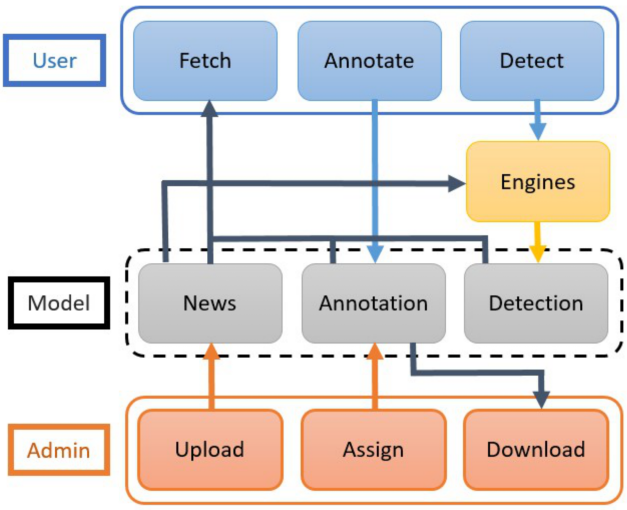

Fig. 1: System Architecture.

Even though the existing automatic fact-checking methods have shown promising results, developing a fully automated fact-checking system is still impractical due to technical difficulties [4] and credibility issues [6]. A more realistic solution to address the scalability issue of manual fact-checking will be to develop an AI-based system to assist human fact-checkers during the fact-checking process. Until recently, very limited work [6], [7] has begun to explore this direction. However, all of them are focused on English news. Our research focuses on fact-checking Cantonese news on Hong Kong local discussion forums. In this paper, we first proposed an AI-based system with new annotation procedures and schema to assist fact-checkers in labeling news on local Hong Kong forums. We then deployed our system and conducted a real case study to demonstrate the effectiveness of our system.

[1]https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory

II. The Proposed Annotation System

Figure 1 illustrates our system with arrows indicating the data flows. The blue (top) and orange (bottom) modules are the graphical user interfaces (GUIs) for users and administrators (admins) respectively. Blocks in the modules are the corresponding functionalities described from bottom to top as follows: (1) **Upload**. Admins upload web-crawled, pre-processed news items in CSV format to the system. The system provides a columns-mapping feature to prevent admins from inducing data errors; (2) **Assign**. Admins select an existing human fact-checker and assign a number of news items for annotation. The assignment creates unique annotation objects with respect to assigned news items and the selected user; (3) **Download**. Admins download completed annotations for downstream analysis; (4) **Fetch**. This function runs upon a human fact-checker login. It fetches the assigned news contents, annotations, and detection outcomes (if available) for GUI display; (4) **Annotate**. The annotation process involves news filtering and news annotation in the form of questions and answers. The user (i.e., fact-checker) sends the annotated data (answers) to the server upon completion. If the detection outcomes are available, the system can ease the process by providing predicted results from automatic detection and highlighting the relevant texts; (5) **Detect**. In case the detection outcomes are not already available for a particular news item, the detection process (the yellow block in Figure 1) automatically starts on the server for automatic detection before the annotation process begins.

The key innovations of our proposed system lie in two functionalities: **Annotate** and **Detect**. Most existing studies on fake news annotation only provide a single final label (true or false). However, fact-checking is a non-trivial interrogate intellectual process, where the exclusive true or false on a given news article will not be enough to reveal the whole reasoning of human fact-checkers. Moreover, without clear annotation guidelines for fact-checking will make inconsistent annotations among different fact-checkers. Through comprehensive discussion with the experienced fact-checkers in our fact-checking center, we propose the following annotation procedure and schema for fact-checking. The procedure contains three steps:

TABLE I: The Proposed Annotation Process and Schema

Annotation Step	Questions
Step 1: News Filtering	Q1: Is the news related to our selected topic? (Yes or No) Q2: Is the news a factual claim? (Yes or No)
Step 2: Pre-Fact Checking	Q3: The source of the news: (A dropdown list) Q4: Incivility of the news: (A dropdown list) Q5: Sentiment of the news: (A dropdown list) Q6: Bias of the news: (A dropdown list) Q7: Worth to check?: (Yes or No) Q8: If yes to Q7, which sentences in the news need to be checked?
Step 3: Fact Checking	Q9: Fact-Checking Decision: (True, Partially True, False or Misleading)

(1) News filtering; (2) Pre-Fact-Checking; and (3) Fact-Checking. In each step, human fact-checkers will follow the schema to answer the designed questions, which assist the reasoning width and depth. The designed questions are shown in Table I. We hypothesize that the answers to Q3 and Q6 in the Pre-Fact-Checking step are indicative features to predict the answer to Q7. If we can correctly predict the answer to Q7, then with the help of our **Detect** functionality, our proposed system can effectively filter the worth-to-check news and save efforts for fact-checkers.

The purpose of **Detect** functionality is to ease the labeling process by providing prediction results from automatic detection models. For the Q4, Q5 in Table I, we have integrated automatic detection models to facilitate the labeling process by providing the prediction results and highlighting the relevant text in the news article. Each news content will pass through three different detection engines. Incivility detection is based on both a self-compiled dictionary and also leverages Google's Perspective API [8]. sentiment analysis is based on a fine-tuned multilingual BERT (mBERT) model.

III. User Interface

In this section, we demonstrate the GUI for the annotation process. Figure 2 shows a two columns layout, where the left is a list of news titles, and the right shows a news panel followed by a dynamic form.

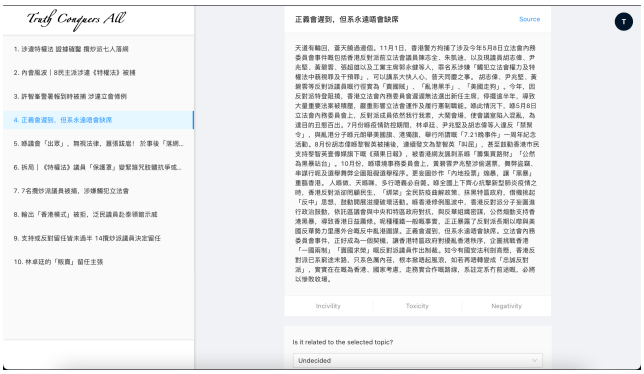

Fig. 2: UI Overview.

News Panel. This area displays essential news information for the annotation process. The top news title is accompanied by a clickable source link on the right, followed by the news content with three detection buttons at the bottom. Clicking on one of the detection buttons displays relevant text highlights and the corresponding prediction results as illustrated in Figure 3.

Dynamic Form. This rule-based form smartly eliminates unwanted news items to minimize the workload. The form comprises a number of key questions and the respective Boolean answers will determine whether the news is relevant for the next step. Figure 4 shows the two outcomes of answering one of the key questions, one answer extends the form with more questions, while the other does not.

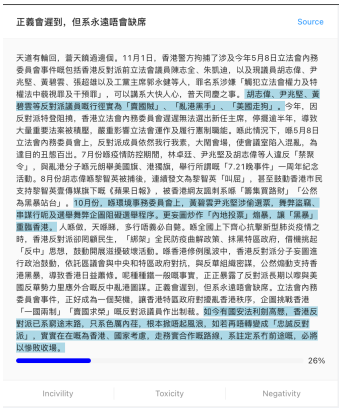

Fig. 3: News Panel.

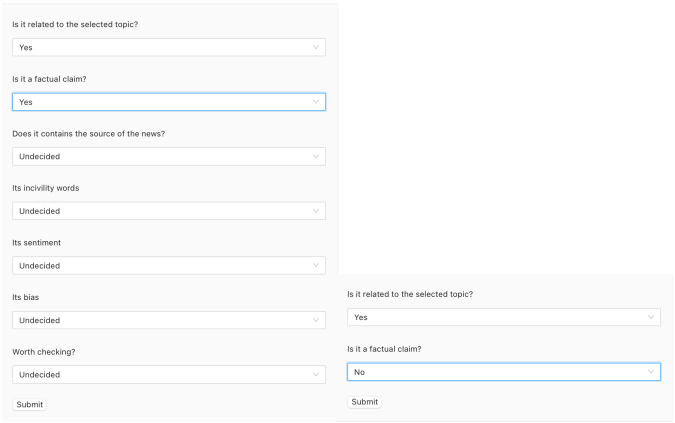

Fig. 4: Dynamic Form.

IV. Real Case Study and Experimental Results

We deployed our system to label and analysis Cantonese fake news related to Anti-Extradition Law Amendment Bill (Anti-ELAB) movement in Hong Kong discussion forums. We have labeled 2,867 posts.

To verify our hypothesis that the answers to Q3-Q6 in the Pre-Fact-Checking step are indicative features to predict the answer to Q7. We have evaluated the performance of building predictive models using the answers from Q3 to Q6 to predict the answer to Q7. Among our labeled posts, there are 2405 no and 462 yes answers to Q7. We use Area under the ROC Curve (AUC) to evaluate the performance. We have tested four popular machine learning classifiers: multiple layer perception (MLP), random forest; Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB). Other than using Q3 to Q6 as the input feature, we also test the performance of these machine learning algorithms by using Term Frequency-Inverse Document Frequency (TF-IDF) of the posts as the input features to predict the answer to Q7.

The results from different machine learning algorithms are shown in Table II. It clearly shows that the designated features (Q3-Q6) achieve better performance than using the TF-IDF representation of news as features. It demonstrates that our proposed annotation system with automatic detection can assist the fact-checker in identifying worth-to-check news effectively.

TABLE II: Prediction Accuracy on Anti-ELAB

Features	MLP	random forest	SVM	XGB
TFIDF	0.59	0.53	0.54	0.59
Q3 - Q6	0.69	0.70	0.68	0.69

V. Conclusion

In this study, we proposed an AI-based system to assist human fact-checkers for labeling Cantonese fake news and deployed it for a real-case study. Using the designed *Annotate* schema, and with the assistance of *Detect* functionality, our proposed system can effectively filter the worth-to-check news and save efforts for fact-checkers.

Acknowledgment

The research was supported by a General Research Fund grant (HKBU 12605520) of the Research Grant Council of Hong Kong, the Public Policy Research Funding Scheme (2021.A2.047.21B) from Policy Innovation and Co-ordination Office of the Hong Kong Special Administrative Region Government, and the Interdisciplinary Research Clusters Matching Scheme (IRCMS/19-20/D04), the Initiation Grant for Faculty Niche Research Areas (RC-FNRA-IG/21-22/COMF/01) of HKBU, and the AIS Scheme (AIS 21-22/01) and CMC Scheme (CMC/21-22/003) from School of Communication, HKBU, and the Natural Science Foundation Council of China (61906161).

References

[1] 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.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[3] 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.

[4] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1803–1812.

[5] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[6] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino, "Automated fact-checking for assisting human fact-checkers," *arXiv preprint arXiv:2103.07769*, 2021.

[7] S. Shaar, F. Alam, G. D. S. Martino, and P. Nakov, "Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document," *arXiv preprint arXiv:2109.07410*, 2021.

[8] A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, "A new generation of perspective api: Efficient multilingual character-level transformers," *arXiv preprint arXiv:2202.11176*, 2022.