MCFEND: A Multi-source Benchmark Dataset for Chinese Fake News Detection

Yupeng Li
Department of Interactive Media
Hong Kong Baptist University
Hong Kong, China
ivanypli@gmail.com

Haorui He
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
hehaorui11@gmail.com

Jin Bai
Department of Computer Science
Beijing Normal University-Hong Kong Baptist University
United International College
Zhuhai, China
jinbai@uic.edu.cn

Dacheng Wen
Department of Computer Science
The University of Hong Kong
Hong Kong, China
wdacheng@connect.hku.hk

ABSTRACT

The prevalence of fake news across various online sources has had a significant influence on the public. Existing Chinese fake news detection datasets are limited to news sourced solely from Weibo. However, fake news originating from multiple sources exhibits diversity in various aspects, including its content and social context. Methods trained on purely one single news source can hardly be applicable to real-world scenarios. Our pilot experiment demonstrates that the F1 score of the state-of-the-art method that learns from a large Chinese fake news detection dataset, *Weibo-21*, drops significantly from 0.943 to 0.470 when the test data is changed to multi-source news data, failing to identify more than one-third of the multi-source fake news. To address this limitation, we constructed the first multi-source benchmark dataset for Chinese fake news detection, termed MCFEND, which is composed of news we collected from diverse sources such as social platforms, messaging apps, and traditional online news outlets. Notably, such news has been fact-checked by 14 authoritative fact-checking agencies worldwide. In addition, various existing Chinese fake news detection methods are thoroughly evaluated on our proposed dataset in *cross-source*, *multi-source*, and *unseen source* ways. MCFEND, as a benchmark dataset, aims to advance Chinese fake news detection approaches in real-world scenarios.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **General and reference** → **Evaluation**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Chinese Fake News Detection, Multi-source Benchmark Dataset, Cross-source Evaluation, Multi-source Evaluation

1 INTRODUCTION

It has been prevalent for the public to consume news through various online sources, such as social platforms and news websites. Such sources are efficient media for spreading fake news. For instance, the latest Weibo annual report on fake news [38] revealed that Weibo's official fact-checking agency identified 82,274 pieces of fake news in 2022. Given the devastating consequences of fake news on both individuals and society, fake news detection has become an urgent and essential task that needs to be addressed [1, 9, 18, 26–28, 36]. To this end, Chinese fake news detection datasets have been constructed for the development of Chinese fake news detection [9–12, 18, 40, 41].

The existing Chinese fake news detection datasets are limited to one single source, Weibo, where both true and fake news are collected. However, in the real world, news emerges from multiple sources, such as social platforms, messaging apps, and traditional online news outlets, etc. [3, 9, 13, 14, 20]. News, in particular, fake news, from different sources is characterized by diverse dimensions, such as news content, topics, publishing methods, and the utilization of sophisticated linguistic styles intended to mimic real news [1, 19, 21, 31, 37]. For example, Fig. 1 shows four instances of fake news, each sourced from a distinct news source, exemplifying different news characteristics. Existing Weibo-based Chinese fake

Figure 1: An example of four pieces of fake news from four different Chinese news sources, including Weibo (a popular social platform), China Times (an online news outlet), Wechat (a messaging app), and Douyin (a social platform). Each piece of fake news showcases different characteristics across various aspects, such as content, topics, publishing methods, linguistic styles, etc.

news detection datasets that fail to capture the above data diversity can impede the effectiveness of machine learning (ML) based fake news detection in practice, including but not limited to the robustness to intricately crafted fake news and the generalization to fake news from other sources [1, 9, 18, 21, 34].

Pilot Experiment. To verify such limitations, we conducted evaluations on 817 pieces of fake news we collected. They were verified between Jan. 2015 and Mar. 2023, from the China Internet Joint Rumor Refuting Platform[1], a government-backed fact-checking agency supported by authoritative experts and various government departments. The agency covers fake news originating from a wide variety of sources, including but not limited to, Douyin, Wechat, TouTiao, Zhihu, Weibo, etc.[2] We trained the state-of-the-art fake news detection model BERT-EMO [41] on the *Weibo-21* dataset [18]. The model demonstrated strong performance with F1 scores of 0.943 on the *Weibo-21*, 0.932 on the *Weibo-20* [41] dataset, and 0.908 on the *Weibo-16* dataset [11], respectively. Nevertheless, when we used the same model to detect fake news collected from all the diverse sources on the platform, failed to identify 35.34% of fake news. The macro F1 score dropped to 0.470, a decline of 52.03%. The following may account for such experimental results. First, during the training phase, ML models lack exposure to a diverse range of data from various sources, which leads to an overfitting of specific characteristics of Weibo fake news. This limits their capability to generalize and effectively identify fake news from different sources. Further, during the testing phase, these models are evaluated using Weibo data exclusively, overlooking a comprehensive assessment of their performance across different news sources. When faced with real-world fake news that emerges from multiple sources, the applicability of models trained

and tested on existing datasets is questionable. Our analytical results validate the limitations of current Chinese fake news detection datasets. Therefore, it is imperative to construct a comprehensive dataset that consists of (real and fake) from diverse sources.

To bridge this gap, we constructed the first M̲ulti-source benchmark dataset for C̲hinese F̲ak̲E̲ N̲ews D̲etection (MCFEND), which contains 23,974 pieces of authoritatively verified Chinese news from *14 fact-checking agencies covering numerous news sources.*[3] These fact-checking agencies are further divided into three distinct groups. The first group encompasses nine Chinese fact-checking agencies. That means the collected news has been verified by experts as active and authoritative. The second group corresponds to three existing annotated English fake news detection datasets. Specifically, for an English news piece paired with its corresponding authenticity label from an existing English fake news dataset, we employ the state-of-the-art cross-lingual identical news retrieval system to collect its Chinese equivalent while retaining its original label. The third group consists of Weibo's official fact-checking agency exclusively, i.e., the Weibo Community Management Center. For this group of news, we utilized news data from the *Weibo-21* dataset [18]. Furthermore, we conducted comprehensive evaluations on six established baseline models for Chinese fake news detection, including state-of-the-art methods, under cross-source and multi-source scenarios on our MCFEND dataset. The experimental results characterize the challenge of accurately spotting fake news from different sources that the dataset presents.

Our contributions are summarized as follows: First, We constructed the initial multi-source Chinese fake news detection dataset MCFEND, which comprises multi-modal content and social context of 23,974 real-world Chinese news pieces collected from 14 authoritative fact-checking agencies in three distinct groups. Additionally, to the best of our knowledge, MCFEND is the largest open-sourced Chinese fake news detection dataset, being at least 2.63 times larger than any existing ones. The dataset aims to benchmark the evaluations of Chinese fake news detection methods in real-world scenarios, where news originates from diverse sources, and to encourage further research in this field. Second, we conducted comprehensive cross-source, multi-source, and unseen source evaluations on six established baseline models for Chinese fake news detection, including state-of-the-art methods. Our experimental results reveal that the models trained on existing datasets are not applicable in real-world scenarios. Incorporating multi-source data is necessary, which can enhance the models' robustness substantially.

2 PRELIMINARIES AND RELATED WORK

Fake news detection, also referred to as false news detection, and related to information credibility evaluation, is commonly defined as a binary classification task [3, 9, 20, 21]. In this context, we define the output space as $\mathcal{Y} = \{0, 1\}$ to classify news items as real (0) or fake (1). The input space $\mathcal{X}$ includes multidimensional data, comprising both the content of the news and its social context. Let

---

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ represent a collection of n news pieces, each associated with a label indicating its authenticity. Here, $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The goal is to design fake news detection models capable of learning a function $\phi : \mathcal{X} \to \mathcal{Y}$, such that for any given news item x_i, $\phi(x_i)$ predicts its label y_i with high accuracy.

Numerous datasets have been constructed to address fake news detection. Representative English fake news detection datasets, such as *BuzzFace* [25], *LIAR* [35], *FakeNewsNet* [30], *PHEME* [43], *KaggleFakeNews* [24], *FakeNewsCorpus* [19], and *FakeHealth* [6], were constructed on English news from social platforms like Twitter and Facebook, as well as fact-checking websites such as BuzzFeed, PolitiFact, and NewsGuard.[4] A few Chinese fake news detection datasets have also been proposed. For instance, Ma et al. introduced the *Weibo-16* dataset [11], collected from the Chinese social platform Weibo. This dataset contains verified fake news sourced from the Weibo Community Management Center[5], an official fact-checking agency for posts on Weibo. Real news was collected from regular posts that were not categorized as fake. While *Weibo-16* focuses exclusively on textual data, Jin et al. [12] later introduced *Media-Weibo*, the first multi-modal dataset for detecting Chinese fake news. *Media-Weibo* includes textual content, user profiles, and supplementary images for each post. Zhang et al. [41] then extended *Media-Weibo* dataset to *Weibo-20* dataset by adding 850 real news pieces authenticated by NewsVerify[6], a fact-checking website dedicated to verifying posts on Weibo, from April 2014 to November 2018, and 1,806 fake news pieces that were officially verified by the Weibo Community Management Center within the same timeframe. Using a similar approach, Yang et al. [40] constructed *CHECKED* dataset, aiming at detecting COVID-19-related fake news on Weibo. Additionally, Nan et al. [18] proposed the *Weibo-21* dataset, the first multi-domain Chinese fake news detection dataset. *Weibo-21* contains both fake and real news pieces collected from Weibo spanning from December 2014 to March 2021, covering nine different domains, such as Science, Military, and Education. Most recently, Hu et al. [10] constructed the multi-modal retrieval augmented dataset *MR2*. This dataset consists of two subsets from Weibo and Twitter, respectively, covering news with images and texts, and provides evidence retrieved from the Internet for both modalities.

Existing datasets for Chinese fake news detection rely heavily on Weibo. Different from them, we constructed the pioneering multi-source Chinese fake news detection dataset, termed MCFEND, which contains 23,974 real-world Chinese news pieces collected from multiple sources across three distinct categories. Table 1 compares the Chinese fake news detection datasets.

3 MCFEND DATASET

In this section, we introduce our Chinese multi-source fake news detection dataset, MCFEND. Additionally, we perform data analysis to investigate into the differences between various sources.

Table 1: Summary of Chinese fake news detection datasets. Please note that the *MR2* dataset comprises two subsets: one from Weibo (Chinese) and one from Twitter (English). The statistics presented in this table specifically pertain to the Weibo (Chinese) subset.

Dataset	#News	Feature			News Source
		Text	Image	Social Context	
Weibo-16	5,656	✓		✓	Weibo
Weibo-Media	5,802	✓	✓	✓	Weibo
Weibo-20	6,362	✓	✓	✓	Weibo
Weibo-21	9,128	✓	✓	✓	Weibo
MR2	6,976	✓	✓	✓	Weibo
MCFEND	23,974	✓	✓	✓	Multiple Sources

3.1 Overview

The MCFEND dataset contains news verified by 14 fact-checking agencies from a wide range of sources, such as messaging apps, social platforms, and traditional news outlets. As mentioned above, these agencies are categorized into three groups (see Table 2. The websites for the 14 fact-checking agencies are detailed in Appendix A. The first group includes nine Chinese fact-checking agencies identified as active by Duke Reporters[7], along with four Chinese fact-checking agencies manually verified by experts as active and authoritative. The second group corresponds to four English fact-checking agencies, including Politifact, Gossipcop, BS Detector, and FakeNewsCorpus. This group contains the Chinese counterparts of the English news fact-checked by the aforementioned agencies, which are collected through a carefully designed cross-lingual identical news retrieval method. An in-depth description of the method can be found in Sec. 3.2.2. Group 3 exclusively covers the Weibo Community Management Center, from which news data was directly sourced from *Weibo-21* dataset [18].

Table 2: An overview of fact-checking agencies.

Group	Fact-checking Agencies
Group 1	China Internet Joint Rumor Refuting Platform
	Tencent Jiaozhen
	China Daily Factcheck
	Taiwan FactCheck Center
	MyGoPen
	HKBU Factcheck
	HKU Annie Lab
	AFP Fact Check Asia
	Factcheck Lab
Group 2	Politifact
	Gossipcop
	BS Detector
	FakeNewsCorpus
Group 3	Weibo Community Management Center

The overall MCFEND dataset contains 23,974 news pieces, including 8,144 sourced from the nine fact-checking agencies in Group 1, 6,702 related to the four English fact-checking agencies in Group 2, and 9,128 obtained from the Weibo Community Management Center in Group 3. Similar to existing dataset construction [18, 30], we collected the following information for each piece of news in any group: (1) Multi-modal news content, including text, images, and metadata, e.g., timestamps; (2) Multi-modal social context, including posts, comments, emojis, user profiles, and other metadata, e.g., like counts of comments. Table 3 presents the detailed statistics of the MCFEND dataset.

Table 3: Statistics of the MCFEND dataset.

Statistics	Group 1	Group 2	Group 3	Overall
#Total	8,144	6,702	9,128	23,974
#Fake	7,486	5,741	4,488	17,715
#Real	473	9,61	4,640	6,074
#user	235,215	156,862	458,800	803,779
#posts	58,299	41,600	70,814	170,713
#comments	262,342	328,465	1,512,095	2,102,902
	Mar. 2015	Jan. 2015	Dec. 2014	Jan. 2015
Timeframe	-	-	-	-
	Mar. 2023	Mar. 2023	Mar. 2021	Mar. 2023

3.2 Dataset Construction

In this subsection, we present the process of constructing the dataset for each group of the fact-checking agencies. Fig. 2 illustrates the entire process for the dataset construction.

3.2.1 Group 1: Fact-checking Agencies Data Crawling. Fact-checking agencies serve as a common source for labeling fake news detection datasets [9, 21]. These agencies are typically operated by government entities, companies, or non-profit organizations, and they employ authoritative experts to assess the authenticity of news pieces originating from diverse sources, such as social platforms, messaging apps, and traditional online news outlets. As discussed in Sec. 1, including a wider range of fact-checking agencies enhance the diversity of news sources in our dataset.

To maximize our coverage of news sources, we conducted web crawling to collect data from all active Chinese fact-checking agencies, encompassing the five Chinese fact-checking agencies identified as active by Duke Reporters, in addition to nine other Chinese fact-checking agencies that were manually verified as active and authoritative. In the case where the labels on some fact-checking agencies, e.g., AFP Fact Check Asia and Factcheck Lab, are presented in the form of images, we utilized an optical character recognition method called Tesseract-OCR to retrieve such labels.[8]

3.2.2 Group 2: Cross-lingual Identical News Retrieval. To further diversify our news sources, we employ a cross-lingual identical news retrieval method to obtain the corresponding Chinese equivalents of both real and fake English news within several well-known datasets, including *FakeNewsNet* [30], *KaggleFakeNews* [24], and

FakeNewsCorpus [19]. *FakeNewsNet* consists of two detailed subcollections sourced from distinct fact-checking organizations, specifically Politifact and Gossipcop. *KaggleFakeNews* contains news gathered from 244 sources classified as "unreliable or otherwise questionable" by the BS Detector, a browser extension that assesses the reliability of websites by comparing them to a professionally curated list. *FakeNewsCorpus* is a dataset consisting of news related to the 2016 US elections. The news pieces in this dataset are manually annotated by its authors.

We consider the BS Detector and the authors of the *FakeNewsCorpus* as two distinct fact-checking agencies. By incorporating these three datasets, we effectively introduce data from four additional fact-checking agencies, enabling us to collect news from a wider range of English news sources.

For each news piece in these datasets, we executed the following steps to identify its corresponding Chinese counterpart:

- Step 1: Translation. We utilized the Baidu Translation API to translate the headlines of the English news into Chinese.[9]
- Step 2: Chinese News Retrieval with Google News.[10] Google News provides extensive and up-to-date news coverage from sources worldwide. We configured the language and region of interest as "Chinese (China)" and employed the translated Chinese news headline as the search query. Search engines typically sort results by relevance. We assumed that the top five returned news pieces were the most relevant Chinese counterparts to the original English news. Subsequently, we crawled the top five returned news pieces.
- Step 3: Cross-lingual News Similarity Calculation. To determine the degree of similarity between the Chinese news retrieved in the previous step and the original English news, we employed the state-of-the-art cross-lingual news similarity calculation system [39], which ranked 1st in the SemEval2022 Task 8 challenge [4] with a Pearson correlation coefficient of 0.818 on the official evaluation set. Specifically, we calculated the similarity score between the retrieved Chinese news and the original English news. The Chinese news with the highest similarity score was preserved in our MCFEND dataset. The strong performance of the system effectively ensures the consistency of misleading content across different languages.
- Step 4: Label Assignment. The authenticity label of the original English news is used to label its Chinese counterpart, that is, the Chinese news with the highest similarity score.

The method inherently retains human-written news content. In contrast to directly utilizing Chinese news content generated by machine translator, our approach avoids unnatural textual expressions that could potentially introduce noise to the models.

3.2.3 Group 3: Weibo News Collection. Group 3 consists solely of news sourced from the Weibo. For this group, we directly utilized news data in the *Weibo-21* dataset [18], the largest Chinese fake news detection dataset on Weibo.

3.2.4 Social Context Collection. Relying solely on news content may be inadequate for detecting fake news, as fake news content

---

[8]https://github.com/tesseract-ocr/tesseract

[9]http://fanyi-api.baidu.com/
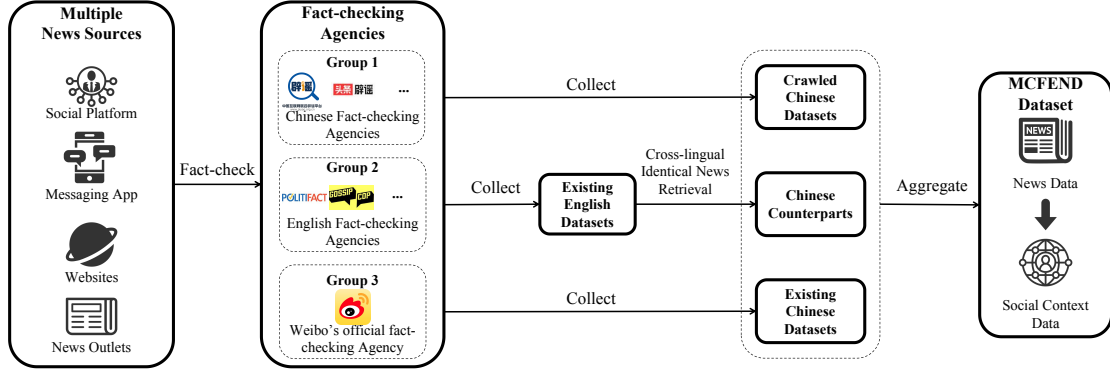[10]https://news.google.com/

Figure 2: The process for constructing the MCFEND dataset.

is often meticulously crafted to deceive the public. Social platforms offer an invaluable source of supplementary information in the form of social context features [9, 16, 17, 29, 30, 32, 41], capturing user interactions and social behaviors within the social platform environment. Thus, to incorporate such important features, we collected social context data, such as posts, comments, user profiles, etc., on the largest social platform in China, Weibo.[11]

The process of collecting social context aligns closely with the approach detailed in [30] for gathering social context from Twitter. Firstly, for news pieces that have headlines, we created search queries for associated posts on Weibo using the headlines. For news pieces without headlines, we utilized the Jieba tool to tokenize the textual content of the news and extract the top five keywords, which were then used as search queries.[12] During this process, we removed special characters from the search queries to eliminate unnecessary noise. All matching posts were considered relevant and included. We then retrieved user responses to these posts, including comments, reposts, and likes. Additionally, upon identifying all users involved in the news propagation process, we collected metadata for these users, such as their usernames and profiles. As shown in Table 3, we assembled a comprehensive set of relevant social context data, which includes 170,713 posts and 2,102,902 comments from 803,779 distinct users.[13]

3.2.5 Post-collection Processing. After collecting all news pieces and their corresponding social context, we conducted three post-collection processing steps:

- Step 1: Text Cleaning. To enhance data quality and eliminate unnecessary noise, we conducted text cleaning on text within both news content and social context. This cleaning process involved removing HTML tags, punctuation, white spaces, stop words, and prefix headings.

- Step 2: Deduplication. The raw data contained multiple duplications. As a result, we removed redundant news and social context data to avoid unnecessary repetitions.
- Step 3: Label Mapping. Different fact-checking agencies employ diverse fine-grained labels to express degrees of authenticity (e.g., true, mostly true, and inconclusive). To ensure consistency, we designed a label mapping strategy to standardize the original labels. Please refer to Appendix B for details of our label mapping strategy.

3.3 Comparison of the Three Groups

Our MCFEND dataset gathers news from 14 fact-checking agencies, categorized into three distinct groups. As previously noted, fake news originating from diverse sources differs in both content and social context [1, 19, 21, 31, 37]. To quantify these differences between the three groups, we conducted analyses on these two aspects, which are considered significant clues for fake news detection [9, 21].

(a) Textual feature.   **(b) Social emotion feature.**
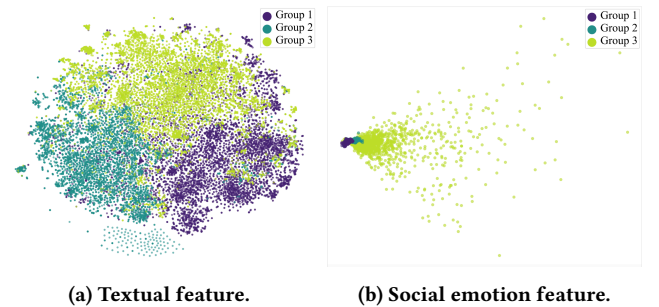
Figure 3: Visualization of textual and social emotion features for news collected from three distinct groups of fact-checking agencies.

Regarding content analysis, we employed a pretrained Chinese Sentence-BERT model [23] to generate text representations for each news piece. Consequently, each news piece was represented as a

---

[11]While Weibo serves as the social context source for all collected news pieces, it also acts as an independent news source.

[12]https://github.com/fxsjy/jieba

[13]Individual users may engage in the social context of news collected from fact-checking agencies across different groups.

768-dimensional vector based on its textual information. This representation serves as an encompassing approximation of the textual features of the news. To visually depict the differences, we applied t-SNE [33] to reduce the dimensionality of these vectors to two.[14] As shown in Fig. 3 (a), the textual features of news pieces collected from fact-checking agencies are distinct. This result validates the significant difference in their textual characteristics.

For the social context, we considered the significant social emotion features [41], which is a 275-dimensional vector representing the emotions evoked within the social context surrounding the news pieces and has been verified to exhibit distinctions between fake and real news [41]. Similar to the analysis above, we reduced the dimensionality of the social emotion feature to two. As illustrated in Fig. 3 (b), the social emotion features associated with news pieces from Weibo demonstrate a more scattered pattern, whereas those from other sources are clustered. These results illustrate the substantial differences in their social context characteristics.

In summary, our analysis of both content and social context provides strong evidence of differences between the three groups. We also consider other factors like text length and dual emotions between the news publisher and associated users in the social context. Due to space limitations, we omit the analysis here.

4 EXPERIMENTS

We conducted experiments to evaluate the performance of representative fake news detection methods on our newly proposed MCFEND dataset. Specifically, we aim to answer the following evaluation questions (EQs):

EQ1 : Are existing methods, which have demonstrated effectiveness on the existing Weibo datasets, capable of maintaining their performance when being applied to news collected from different sources?

EQ2 : Can training with multi-source data enhance the robustness of existing methods in detecting fake news in real-world scenarios, which involve multiple sources?

EQ3 : Can training with multi-source data enhance the robustness of existing methods in detecting fake news originating from previously unseen news sources?

The detailed experimental setups are provided in Appendix C.

4.1 Baselines

To address the above EQs, we carefully selected six baseline models spanning the two widely recognized categories of fake news detection approaches: content-based and social context-based methods [1, 9, 10, 42], and constructed the baseline benchmarks. Their implementation details are as follows.

4.1.1 Content-based Methods. Content-based methods rely on solely the textual or visual contents of the news. We adopted two representative types of content-based models, *uni-modal models* and *multi-modal models*.

Uni-modal models focus on the textual content of the news. We used BERT [7] and RoBERTa [15] as contextualized encoders to encode the textual content. Then, the representation of the "[CLS]"

special token is used for prediction. The implementation of BERT and RoBERTa in this study is based on their respective Chinese base versions, i.e., BERT-base-Chinese[15] and RoBERTa-wwm-Base.[16] *Multi-modal models* encode both text and images in the input news. We consider two multi-modal baselines: CLIP [22] and CAFE [5]. CLIP [22] is a pretrained model for images and text. We input the image and text of the news into CLIP to obtain a joint representation of the visual and textual content. This joint representation is then utilized for predictions. We implement CLIP based on its Chinese version.[17] CAFE [5] is an ambiguity-aware fake news detection method. Specifically, it integrates uni-modal features produced by BERT [7] for text and ResNet [8] for images, along with cross-modal correlations. It relies on uni-modal features when cross-modal ambiguity is weak and relies on cross-modal correlations when cross-modal ambiguity is strong. CAFE demonstrates superior fake news detection performance on the *Twitter* [2] and *Weibo-16* [11] datasets, respectively. It stands as the state-of-the-art content-based approach to this task.

4.1.2 Social Context-based Methods. Social context-based methods are typically categorized into three groups: tree-based, modal fusion-based, and graph-based [10]. Since our MCFEND dataset contains news from diverse sources and lacks the necessary cross-source user/news interactions to build effective graphs and trees, we included only the *modal fusion-based models*, excluding the graph-based and tree-based models.

Modal fusion-based models integrate information from both news content and social context. In our study, we considered two representative baseline models for this category: dEFEND [29] and BERT-EMO [41]. The dEFEND model [29] utilizes a sentence-comment co-attention sub-network to exploit news contents and comments in the social context to jointly capture explainable top-k check-worthy sentences and comments for fake news detection. On the other hand, BERT-EMO [41] enhances a BERT-based fake news detector by incorporating dual emotion features that represent both the emotions and the relationship between news and comments within the social context. Note that the BERT-EMO model has demonstrated outstanding performance in fake news detection, achieving the highest reported performance on the *Weibo-20* dataset [41]. Besides, our preliminary experiments on the *Weibo-21* dataset [18] have also shown that BERT-EMO achieved an impressive F1-score of 0.943, outperforming all other methods. The results show that BERT-EMO as the state-of-the-art social context-based approach in the fake news detection task.

4.2 Cross-source Evaluation

To address EQ1, we performed cross-source evaluations on the baseline models. Specifically, we evaluated the performance of baseline models trained exclusively on Weibo data, focusing on their application to diverse news sources. Our findings, shown in Table 4, reveal significant variations in performance across different test groups, underscoring the challenge of applying models trained on a single-source dataset to news originating from varied sources.

---

[14]Other dimensionality reduction techniques like PCA and UMAP are applicable, but due to space limits, we only present the results of t-SNE here.

[15]https://huggingface.co/bert-base-chinese
[16]https://huggingface.co/hfl/chinese-roberta-wwm-ext
[17]https://github.com/OFA-Sys/Chinese-CLIP

Table 4: The performance of the baselines in cross-source evaluation. The training data for all baselines in cross-source evaluation is exclusively sourced from Weibo. The overall results are calculated using test data from all groups. The highest number in each group is in bold.

Model	Test Data	Accuracy	Macro F1
BERT	Group 1	**0.830±0.018**	**0.521±0.009**
	Group 2	0.453±0.027	0.407±0.017
	Group 3	0.817±0.006	0.816±0.006
	Overall	0.719±0.008	0.673±0.007
RoBERTa	Group 1	0.124±0.008	0.123±0.007
	Group 2	**0.763±0.038**	**0.539±0.016**
	Group 3	0.883±0.006	0.883±0.006
	Overall	0.595±0.011	0.582±0.009
CLIP	Group 1	0.769±0.011	0.503±0.006
	Group 2	0.500±0.043	0.458±0.030
	Group 3	0.895±0.012	0.895±0.012
	Overall	0.741±0.009	0.711±0.009
CAFE	Group 1	0.586±0.013	0.431±0.006
	Group 2	0.343±0.032	0.335±0.026
	Group 3	0.891±0.001	0.891±0.001
	Overall	0.635±0.014	0.621±0.013
dEFEND	Group 1	0.656±0.031	0.452±0.009
	Group 2	0.511±0.058	0.400±0.026
	Group 3	0.743±0.016	0.743±0.016
	Overall	0.709±0.002	0.701±0.003
BERT-EMO	Group 1	0.572±0.043	0.405±0.013
	Group 2	0.343±0.174	0.287±0.104
	Group 3	**0.943±0.010**	**0.943±0.010**
	Overall	**0.821±0.008**	**0.818±0.008**

The BERT-EMO model delivered the highest performance, achieving an accuracy of 0.821 and a macro F1 score of 0.818. However, while it scored a high macro F1 of 0.943 on data from Weibo (Group 3), its effectiveness significantly decreased on data from Group 1 and Group 2, with scores dropping to 0.572 and 0.343, respectively. Note that the pattern of performance variance was consistent across all baseline models, including BERT, RoBERTa, CLIP, CAFE, and dEFEND, across different groups. *These findings provide a crucial insight in the responses to EQ1: Baseline models, even those considered state-of-the-art and trained on Chinese fake news detection datasets from Weibo, exhibit limited robustness when confronted with fake news from diverse sources in the wild.* One contributing factor to such performance decrease may be the significant difference in the content and social context feature, as shown in Fig. 3, between news sourced from Weibo and news from other sources. This discrepancy also highlights a significant concern regarding the possible overestimation of the effectiveness of current Chinese fake news detection methods, underscoring the necessity for a thorough reevaluation before they are considered for practical application in real-world situations.

A comparison in Table 5 shows the average decrease in macro F1 score when moving from Weibo-sourced data to other sources, with a smaller decrease indicating greater source robustness. Our analysis found that models employing modal fusion approaches,

integrating both text and social context, demonstrate stronger resilience against data source variability, with an average decrease of only 0.084. In particular, the modal fusion-based model, dEFEND, exhibited the greatest robustness, with a minimal decrease of 0.032. In contrast, uni-modal and multi-modal approaches saw larger decreases of 0.222 and 0.227, respectively. These results indicate that the model performance is greatly influenced by content pattern differences among sources, suggesting that exploring ways to mitigate these impacts is a valuable direction for future research.

Table 5: The comparison of the average performance decrease of all baseline categories, which represents the difference between overall performance and performance on data sourced from Weibo. A smaller performance decrease indicates greater cross-source performance robustness.

Baseline Category	Average Macro F1 Decrease
Uni-modal	-0.222
Multi-modal	-0.227
Modal fusion-based	**-0.084**
Average	-0.107

4.3 Multi-source Evaluation

To address EQ2, we performed multi-source evaluations on the selected baseline models. Specifically, these baseline systems were trained with the complete train set of the MCFEND dataset, including news from all sources covered in our dataset and their corresponding social contexts.

The multi-source evaluation results, depicted in Table 6, provide crucial insights. A comparison with the cross-source evaluation data (refer to Table 4) reveals significant performance improvements in all baseline models upon integrating multi-source data for training. Notably, RoBERTa and CAFE show the most substantial gains, with their macro F1 scores increasing by 0.245 and 0.224, respectively. Such boost in performance can be attributed to the diverse range of fake news features presented by multi-source data, which enhances the models' ability to discern between fake and real news from different sources, helping to prevent models from overfitting to the specific characteristics of data from a single source like Weibo. To offer a qualitative analysis of the enhancement from using multi-source data in the training process, we take the CAFE model as an example. When trained with Weibo data exclusively, the CAFE model fails to correctly identify the news pieces (c) and (d) shown in Fig. 1 as fake. However, when trained with data from all the diverse sources encompassed in the MCFEND train set, the CAFE model exhibits the ability to accurately detect all of the presented fake news in Fig. 1.

These findings address EQ2 by demonstrating that using multi-source data to train fake news detection models significantly improves their performance and robustness in real-world applications. MCFEND dataset could be an invaluable asset for enhancing the detection of Chinese fake news across a variety of sources.

Additionally, the overall performance of almost all baseline models, except dEFEND, when they were trained and tested on multi-source data is lower than their performance when they were trained

Table 6: The performance of the baselines in multi-source evaluation. The training data for all baselines in multi-source evaluation contains news from all sources. The overall results are calculated using test data from all groups. The highest number in each group is in bold.

Model	Test Data	Accuracy	Macro F1
BERT	Group 1	**0.934±0.004**	0.535±0.015
	Group 2	0.858±0.007	0.614±0.052
	Group 3	0.803±0.015	0.801±0.015
	Overall	0.862±0.005	0.800±0.012
RoBERTa	Group 1	0.939±0.002	0.518±0.012
	Group 2	**0.860±0.001**	0.501±0.013
	Group 3	0.844±0.011	0.844±0.011
	Overall	0.880±0.004	0.827±0.008
CLIP	Group 1	0.8891±0.020	0.5213±0.018
	Group 2	0.690±0.073	0.565±0.044
	Group 3	0.8518±0.005	0.851±0.005
	Overall	0.8178±0.015	0.7685±0.014
CAFE	Group 1	0.926±0.008	**0.563±0.016**
	Group 2	0.857±0.010	**0.675±0.007**
	Group 3	0.875±0.008	0.875±0.008
	Overall	0.887±0.007	0.845±0.007
dEFEND	Group 1	0.915±0.005	0.521±0.025
	Group 2	0.755±0.067	0.482±0.065
	Group 3	0.757±0.004	0.756±0.004
	Overall	0.791±0.007	0.774±0.007
BERT-EMO	Group 1	0.902±0.004	0.514±0.040
	Group 2	0.827±0.027	0.545±0.052
	Group 3	**0.937±0.006**	**0.937±0.006**
	Overall	**0.922±0.005**	**0.916±0.005**

and tested on Weibo data exclusively. This finding addresses the challenge of developing algorithms capable of effectively distinguishing generic fake news features across news from various sources in real-world scenarios. Compared among baseline models, the state-of-the-art content-based and social context-based models, CAFE and BERT-EMO, stand out with macro F1 scores of 0.845 and 0.916, respectively. This implies that their relatively lower performance in cross-source evaluation suggests limitations in current training datasets. Thus, our diverse and comprehensive MCFEND dataset, designed to address these limitations, is crucial for advancing Chinese fake news detection in the wild.

5 UNSEEN SOURCE EVALUATION

To address EQ3, we conducted an unseen source evaluation to assess the robustness of existing methods in detecting fake news from previously unencountered news sources. This involved training two versions of the BERT-EMO model, which had shown superior performance in both cross-source and multi-source evaluations, on distinct dataset compositions. Model A was trained exclusively on data from Group 3 (Weibo), while Model B incorporated data from both Group 1 (various Chinese news sources verified by fact-checking agencies) and Group 3 (Weibo).

Table 7: Macro F1 Score Comparison on Unseen Sources.

Model	Train Data	Accuracy	Macro F1 Score
Model A	Group 3	0.343±0.174	0.287±0.104
Model B	Groups 1 and 3	**0.602±0.204**	**0.432±0.088**

We then evaluated both models using data from Group 2 (English news sources), which was new to each model. The results, detailed in Table 7, reveal that Model B, trained on a more diverse dataset, achieved a higher macro F1 score of 0.432, compared to Model A's score of 0.287. This finding addresses EQ3 by suggesting that leveraging multi-source data enhances the robustness of methods for detecting fake news from unseen sources. Therefore, when a new news source emerges in the future, but there is no available data from that platform for model training, a model trained on multi-source data can be expected to more accurately detect fake news from this unseen new source.

6 CONCLUSION

In this work, we introduced the first multi-source benchmark dataset for Chinese fake news detection, termed MCFEND. Unlike existing Chinese fake news detection datasets that are based on a single news source, i.e., Weibo, MCFEND is constructed on (real and fake) news from multiple sources that were fact-checked by 14 authoritative fact-checking agencies. To test the applicability of existing methods, we conducted a systematic evaluation of six representative fake news detection models, including the state-of-the-art ones, in both cross-source and multi-source scenarios. Our experimental results reveal that models trained exclusively on Weibo data can hardly be applicable in real-world scenarios, where fake news typically originates from diverse sources. We also found that incorporating multi-source data into model training enhances the robustness of existing fake news detection methods. Our proposed MCFEND aims to be a benchmark dataset for Chinese fake news detection in the wild, which advances new effective methods in this research field.

Our world is increasingly suffering from unrestrained spread of misinformation in many areas. We hope MCFEND can put forward research results that can combat misinformation and help make our world a better one.

7 ACKNOWLEDGEMENT

REFERENCES

[1] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake News, Disinformation and Misinformation in Social Media: A Review. *Social Network Analysis and Mining* 13, 1 (2023), 30.

[2] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and Visualization of Misleading Content on Twitter. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 71–86.

[3] Alessandro Bondielli and Francesco Marcelloni. 2019. A Survey on Fake News and Rumour Detection Techniques. *Information Sciences* 497 (2019), 38–55.

[4] Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 Task 8: Multilingual News Article Similarity. In *Proc. of SemEval*. 1094–1106.

[5] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-Modal Ambiguity Learning for Multimodal Fake News Detection. In *Proc. of WWW*. 2897–2905.

[6] Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger Cannot Cure Cancer: Battling FakeHealth News with a Comprehensive Data Repository. In *Proc. of ICWSM*. 853–862.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*. 4171–4186.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*. 770–778.

[9] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep Learning for Fake News Detection: A Comprehensive Survey. *AI Open* 3 (2022), 133–155.

[10] Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In *Proc. of SIGIR*. 2901–2912.

[11] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proc. of MM*. 795–816.

[12] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proc. of MM*. 795–816.

[13] Yupeng Li, Haorui He, Shaonan Wang, Francis C.M. Lau, and Yunya Song. 2023. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems* (2023).

[14] Yupeng Li, Dacheng Wen, Haorui He, Jianxiong Guo, Xuan Ning, and Francis C. M. Lau. 2023. Contextual Target-Specific Stance Detection on Twitter: Dataset and Method. In *Proc. of IEEE ICDM*. 359–367.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[16] Jing Ma and Wei Gao. 2020. Debunking Rumors on Twitter with Tree Transformer. In *Proc. of COLING*. 5455–5466.

[17] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proc. of ACL*. 1980–1989.

[18] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MD-FEND: Multi-Domain Fake News Detection. In *Proc. of CIKM*. 3343–3347.

[19] Archita Pathak and Rohini Srihari. 2019. BREAKING! Presenting Fake News Corpus for Automated Fact Checking. In *Proc. of ACL-SRW Workshop*. 357–362.

[20] Francesco Pierri and Stefano Ceri. 2019. False News on Social Media: A Data-driven Survey. *ACM Sigmod Record* 48, 2 (2019), 18–27.

[21] Piotr Przybyla. 2020. Capturing the Style of Fake News. In *Proc. of AAAI*. 490–497.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. of ICML*. 8748–8763.

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of EMNLP*.

[24] Megan Risdal. 2016. Getting Real about Fake News. https://doi.org/10.34740/KAGGLE/DSV/911

[25] Giovanni Santia and Jake Williams. 2018. BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos. In *Proc. of ICWSM*. 531–540.

[26] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. In *Proc. of ACL*. 4543–4556.

[27] Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article Reranking by Memory-Enhanced Key Sentence Matching for Detecting Previously Fact-Checked Claims. In *Proc. of ACL*. 5468–5481.

[28] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning. In *Proc. of CIKM*. 1640–1650.

[29] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. DE-FEND: Explainable Fake News Detection. In *Proc. of KDD*. 395–405.

[30] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (2020), 171–188.

[31] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2021. Early Detection of Fake News with Multi-source Weak Social Supervision. In *Proc. of ECML PKDD*. 650–666.

[32] Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proc. of ACL-IJCNLP*. 1556–1566.

[33] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[34] Jindong Wang, Haoliang Li, Haohan Wang, Sinno Jialin Pan, and Xing Xie. 2023. Trustworthy Machine Learning: Robustness, Generalization, and Interpretability. In *Proc. of KDD*. 5827–5828.

[35] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proc. of ACL*. 422–426.

[36] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proc. of KDD*. 849–857.

[37] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak Supervision for Fake News Detection via Reinforcement Learning. In *Proc. of AAAI*. 516–523.

[38] Weibo. 2023. Weibo's Annual Repot on Fake News. https://weibo.com/1866405545/MoICtozcu?type=repost.

[39] Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. HFL at SemEval-2022 Task 8: A Linguistics-inspired Regression Model with Data Augmentation for Multilingual News Similarity. In *Proc. of SemEval*. 1114–1120.

[40] Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. CHECKED: Chinese COVID-19 Fake News Dataset. *Social Network Analysis and Mining* 11, 1 (2021), 58.

[41] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection. In *Proc. of WWW*. 3465–3476.

[42] Yiyuan Zhu, Yongjun Li, Jialiang Wang, Ming Gao, and Jiali Wei. 2024. FaKnow: A Unified Library for Fake News Detection. *arXiv preprint arXiv:2401.16441* (2024).

[43] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting Context for Rumour Detection in Social Media. In *Social Informatics*. 109–123.

A THE FACT-CHECKING AGENCIES

Table 8 lists the websites for the 14 fact-checking agencies.

B LABEL MAPPING STRATEGY

Various fact-checking agencies use distinct labels to indicate the truthfulness of information, such as "rumor", "mistake", and "misleading". To unify the label space in our dataset, we used a mapping strategy to standardize the original labels, as illustrated in Table 9. We also provide the English translations of these labels.

C EXPERIMENTAL SETUP

Consistent with prior studies on detecting fake news, our evaluation framework utilizes accuracy and the macro F1 score as metrics. We divided the MCFEND dataset into training, validation, and test segments following a 7:1.5:1.5 ratio through random division to ensure our experimental results' reliability. This division was conducted three times, with each cycle running for 100 epochs on a single NVIDIA V100 GPU. We report the mean F1 and accuracy scores along with their standard deviations from the three iterations. All models were trained using the Adam optimizer with a batch size of 32. To increase training efficiency, we applied an early stopping rule, halting the training if no performance improvement was seen after 500 consecutive batches, ensuring model convergence.

Hyper-parameter selection was based on the validation set's performance. A comprehensive summary of the hyper-parameter settings for each model is provided in Table 10. For parameters not

covered here, we suggest consulting the original publications for more details. We standardized the preprocessing of text to a maximum of 256 tokens and images to a uniform size of 224x224 pixels through center-cropping. In scenarios where news items did not have associated images, we used a standard blank white image for training multi-modal models.

Table 8: Fact-Checking Agencies.

Fact-Checking Agency	URL
China Internet Joint Rumor Refuting Platform	https://www.piyao.org.cn/
Tencent Jiaozhen	https://vp.fact.qq.com/
China Daily Factcheck	https://www.chinadaily.com.cn/china/factcheck/
Taiwan FactCheck Center	https://tfc-taiwan.org.tw/
MyGoPen	https://www.mygopen.com/
HKBU Factcheck	https://factcheck.hkbu.edu.hk/home/
HKU Annie Lab	https://annielab.org/
AFP Fact Check Asia	https://factcheck.afp.com/afp-asia/
Factcheck Lab	https://www.factchecklab.org/
Politifact	https://www.politifact.com/
Gossipcop	https://www.gossipcop.com/
BS Detector	https://github.com/selfagency/bs-detector
FakeNewsCorpus	https://github.com/architapathak/FakeNewsCorpus
Weibo Community Management Center	https://service.account.weibo.com

Table 9: The strategy for mapping the authenticity labels by the fact-checking agencies to our standardized labels.

Standardized Label	Fact-checking Agency Labels
Fake	谣言 (Rumor), 错误 (Mistake), 误导 (Misleading), 部分错误 (Partial Error), 假 LINE (Fake Line), 易误解 (Easily Misunderstood), 虚假 (Fake), 假影片 (Fake Video), 诈骗 (Scam), 假新闻 (Fake News), 假讯息 (Fake Information), 假知识 (Fake Knowledge), 假图片 (Fake Picture), 假脸书 (Fake Facebook), 假健康 (Fake Health), 误导内容 (Misleading Content), 假养生 (Fake Health Care), 部分真实 (Partial Truth), 改动照片 (Altered Picture), 假食安 (Fake Food Safety), 易误导 (Easily Misleading), 误传 (Misinformation), 假标题 (Fake Title), 假科学 (Fake Science), 假好康 (Fake Good Health), 假医学 (Fake Medicine), 真诈骗 (Real Scam), 假照片 (Fake Photo), 当心 (Be Careful), 变假的 (Become Fake), 是假的 (Fake), 假公益 (Fake Charity), 假寻人 (Fake Person Search), 标题党 (Sensational Headlines), 假美容 (Fake Beauty), 假中毒 (Fake Poisoning), 错误及没有证据 (Error and No Evidence), 没有证据及误导内容 (No Evidence and Misleading Content), 脸书诈骗 (Facebook Scam), 非法 (Illegal), 不实来源 (False Sources), 不实 (False), 部份错误 (Partial Error), 惯犯 (Recidivist), 别去了 (Don't Go), 假民俗 (Fake Folklore), 假通知 (Fake Announcement), 诈骗的 (Fraudulent), 防诈骗 (Be Careful for Scams), 防谣言 (Be Careful for Rumors), 请提防 (Please Be Cautious), 过期片 (Outdated Video), 老谣言 (Old Rumor), 触法 (Illegal), 错误标题 (False Title), 错误及误导内容 (False and Misleading Content), 旧新闻 (Past News), 宣导 (Propaganda).
Real	事实 (Fact), 查证 (Verified), 真活动 (Real Activity), 真影片 (Real Video), 误解 (Misunderstanding), 这真的 (This Is True), 真实 (Real), 解释 (Explain), 政策 (Policy), 还真的 (It's Really True), 非谣言 (Non-Rumor), 公告 (Announcement), 已核实 (Verified), 非诈骗 (Not Fraud), 正确 (Correct), 真图片 (Real Picture), 别误解 (Don't Misunderstand), 真照片 (Real Photo), 真的 (Real), 纯热心 (Pure Enthusiasm), 真公益 (True Charity), 勿恐慌 (Don't Panic), 观点 (Viewpoint), 真研究 (Real Research), 真服务 (Real Service), 活动 (Activity), 防灾 (Disaster Prevention), 无责任 (No Responsibility), 澄清 (Clarification), 勿过量 (Don't Overdose), 勿惊慌 (Don't Panic), 真罚单 (Real Ticket), 证实 (Confirmed).

Table 10: Model-wise core hyper-parameter settings adopted in our implementations of the baselines.

Model	Hyper-parameters
BERT	learning rate: 2e-5; hidden size: 768; number of layers: 12; number of attention heads: 12.
RoBERTa	learning rate: 2e-5; hidden size: 768; number of layers: 12; number of attention heads: 12.
CLIP	learning rate: 2e-5; project_dim: 512.
CAFE	learning rate: 1e-3; text_embedding: fixed BERT embedding; image_embdding: fixed ResNet embedding.
dEFEND	learning rate: 2e-5; text_embedding: fixed BERT embedding; project_dim: 200; number of RNN layers: 2.
BERT-EMO	learning rate: 1e-3; hidden_size (bidirectional-GRU): 32; hidden_size (fully connected layers): 32.