Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Data Annotation Methodologies for Fake News

Ruiyi Wang*

School of Economic and Management, Communication University of China, Beijing, China

*Corresponding author: 202219263001@mails.cuc.edu.cn

Abstract. With the development of technology, information dissemination has become faster and more convenient. Fake news has drawn much attention due to its characteristics, such as rapid spread, strong disguise ability, and great harm. The performance of existing fake news detection models is highly dependent on the quality of training datasets. It is crucial to construct high-quality and lower-cost training datasets. The research progress of fake news dataset construction is systematically reviewed in this paper. Firstly, the categories and definition of fake news and the summary of existing mainstream datasets for detecting fake news are reviewed in this paper. Secondly, for traditional text news and newly derived multimodal news, the advantages and disadvantages of the existing annotation technologies are analyzed starting from the three aspects of traditional manual annotation, semi-automated annotation, and dynamic annotation. Finally, future research directions are proposed to address the problems of current datasets in dynamic annotation, multimodal fusion, and cross-domain generalization. High-quality datasets can effectively promote the development of fake news detection technology to meet the challenges of the increasingly complex network information environment.

Keywords: Fake news; Data annotation; Fake news detection model.

1. Introduction

With the development of technology and the Internet, the possibility of generating fake news has greatly increased. At the same time, the ways of producing and spreading fake news are endless and extremely harmful, and may even endanger social order and hinder social security. Therefore, the timely detection of fake news is particularly important. The use of reinforcement learning to detect early fake news was first proposed to curb its spread in [1]. Considering the advantages of Large Language Models in the field of fake news detection. A detection system based on the adaptive basic principle is designed by integrating the LLM into the Small Language Model (SLM) in [2]. Broad and domain-independent features are used to distinguish fake news from legitimate news in [3]. Considering the detection efficiency of fake news, a fake news detection model based on adaptive knowledge perception is proposed in [4], which improves the detection efficiency of fake news by introducing external knowledge as a supplement to the factual background. In addition, models such as Transformer [5], Bidirectional Encoder Representations from Transformers (BERT) [6], and Convolutional Neural Networks (CNN) [7] have also been widely used to extract the semantics of news text and fuse multimodal features, and have been combined with knowledge graphs and attention mechanisms to detect fake news in various fields effectively. Although numerous models have been sufficiently developed, the lack of universal datasets is attributed to the timeliness of news and the flexibility of features. Optimal model performance is achieved only when the training dataset is sufficiently large, content-rich, and accurately classified.

The construction of fake news datasets will be reviewed in this paper. First, the basic definition of fake news and the types of datasets currently constructed are introduced. Secondly, a summary of the fake news data annotation and dataset construction is given. Finally, future research directions are proposed.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

2. Related Works

2.1. Fake news

Fake news is defined as intentionally fabricated or misleading news reports that contain factual inaccuracies. Fake news is frequently designed to emulate authentic news formats, thereby enhancing perceived credibility and enabling financial gains [8]. The current mainstream methods for labeling fake news data are mainly crowdsourcing and semi-automatic annotation. Crowdsourcing refers to finding a group of volunteers from different fields and skills to complete the same given task. Participants may participate for their benefit or for some external reasons to complete the task. Semi-automatic annotation is based on training an automatic annotation model based on a small number of manually annotated datasets. After the model gives the results, experts will review and modify them. At the same time, a variety of fake news fact-checking platforms have emerged, such as PolitiFact, Buzzfeed, and Snopes[9]. These platforms have collected a large amount of fake news and labeled it, providing the public with a platform to identify fake news.

2.2. Introduction to Mainstream Datasets

The mainstream datasets in the field of fake news detection are summarized in this paper, as shown in Table 1. There are some mainstream datasets in the field of fake news detection. FakeNewsNet mainly includes news articles and their auxiliary information, and focuses on social media such as user metadata and news comments. The Weibo dataset is also constantly updated. In addition to the news content, Weibo-20 also contains external information, such as the number of likes and comments, and user information. Visual information is added to the Weibo-21 dataset. PHEME is available in two languages, Chinese and English, and contains multimodal data.

Table 1. Mainstream datasets

Name	Language	Multimodal	Features	Fake News Amount	Real News Amount
FakeNewsNet	English	No	Article URL, Title, Tweet ID	5775	17441
LIAR	English	No	Authenticity label, Statement Content, Context, Subject Content	6418	6418
Weibo-20	Chinese	No	Likes, comments, reposts, user information, URL (tweets, users)	3161	3201
Weibo-21	Chinese	Yes	Likes, comments, reposts, user information, URL (tweets, users, pictures)	4488	4640

3. Research Detection of fake news dataset annotation methods

3.1. Manual annotation method

Manual labeling methods are divided into expert labeling and crowdsourcing labeling. To address the limitation of data sources, news was collected in pairs through manual verification and crowdsourced quality control in [10]. Fake versions were then generated based on the summary by manually imitating the language style of fake news. At the same time, the differences in language characteristics of fake news across fields were compared. It was found that fake news paid more attention to the use of social vocabulary, while real news paid more attention to cognitive expression. A structured news credibility assessment framework is proposed in [11]. It defined content indicators (title representativeness, external expert citations, logical fallacies) and context indicators (fact checking, advertising features), and discarded useless indicators to build a credibility labeling system that can be shared across platforms. An explainable reliability annotation framework, Reliable and Unreliable

News Annotation Schemes (RUN-AS), was proposed in the [12], which provides reliability based on the news text content itself through fine-grained semantic analysis. This scheme can label the structural parts and basic content elements of news, as well as some interest element labels (headlines, inflammatory remarks, etc.), and assign binary reliability ratings to content labels.

In analyzing news structure and content, the inverted pyramid structure and 5W1H analysis method are often used, as shown in Fig.1. The inverted pyramid structure is a way of presenting news according to the importance, freshness, and interest of the news facts. It includes the title and introduction at the top, the main body in the middle, and the ending at the bottom. The main part is written mainly through the 5W1H. 5W1H stands for Who, What, When, Where, Why, and How. All the content that news reports need to explain and illustrate is covered in this formula. By giving 5W1H labels, the true content of the news can be fully understood, and the key issues when labeling data can be identified, thereby reducing the misjudgment rate of the model.

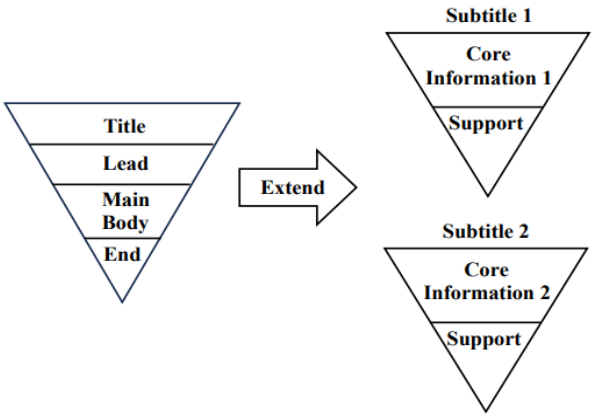

Fig. 1 Inverted pyramid structure and 5W1H structure

3.2. Semi-automatic dataset annotation

3.2.1. AI-assisted annotation

The development of artificial intelligence has become very powerful and can replace humans in doing numerous repetitive tasks. LLM is used for text annotation, and a manual review step is added to ensure the validity of the results in [13]. A semi-automatic annotation method for complex tasks based on the Human-in-the-loop (HITL) paradigm is proposed in [14]. It is divided into three stages for annotation. In the first stage, a small-scale manual annotation was performed according to the RUN-AS guidelines to ensure the quality of the data. In the second stage, active learning methods and HITL strategy are used to integrate automatic and manual annotations to select the most valuable news, and news that the model is uncertain about is manually annotated and reviewed. In the third stage, the data annotated in the first two stages were used to train the news automatic annotation model, and the BETO Transformer model was used to extract more critical information from the core content of the news. Only the annotation results were manually reviewed and modified at the end. The annotation efficiency of the third stage has been greatly improved. An improved dataset construction method is proposed, and the author focused on upgrading the classification system from authenticity rating to reliability rating in [15]. Compared with [14], the dataset of [15] is more concise, and the method based on the text itself is more accurate. Research shows that the “model + manual” annotation method can improve annotation efficiency and reduce costs. At the same time, manual labeling is highly subjective, which can improve the objectivity of labeling to a certain extent and reduce subjective errors.

3.2.2. Semi-supervised and unsupervised annotation

1. Text News

Currently, full automation is not possible. However, semi-supervised and unsupervised annotation only require a small amount of manually labeled data. Machines and models can be used for

subsequent annotation, making it easy to obtain more high-quality annotated data. A cross-stitch-based semi-supervised attention neural model is proposed, and a large-scale annotated dataset is constructed in [16]. To address the problem of the lack of labeled datasets, semi-supervised learning uses the cross-stitch mechanism to obtain the best combination of input model parameters and annotate valuable data. The labeled data was put into Cross-SEAN, and it was found that the model can well identify fake news about COVID-19. A method for automatic news annotation based on verified fact-checking statements on Online Social Networks (OSN) is proposed in [17]. In the data collection section, include journalists' and experts' assessments as supporting statements. Two data crawling methods are followed, without text (only title and basic judgment) and with text (new support statements and news text content). In the data processing part, the labels are unified into two categories. The BERT Transformer model is used in the automatic annotation model to embed the title and text, and the pairwise cosine distance between the title and text is calculated according to the defined threshold. Use conventional majority voting and weighted voting to give the final judgment label:

(1) Regular majority voting: If all three statements are fake, then the news is fake.

(2) Weighted voting: Weighted voting is performed based on the relevance score and cosine similarity of each top statement. Statements with higher relevance scores and similarity have greater weights.

If the model's performance is improved compared with the previous model after training the model using the dataset annotated by the model, it means that the dataset plays an important role for the detector. The model mainly uses indicators such as Precision, Accuracy, Recall, and F1-score to judge its performance:

$$\text{Precision}=\frac{TP}{TP+FP} \tag{1}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

$$\text{Recall}=\frac{TP}{TP+TN} \tag{3}$$

$$F1=\frac{2TP}{2TP+FP+FN} \tag{4}$$

The "Search and Examine and Early-Termination" (SEE) model is proposed in [18]. That is a model that uses data that does not require manual annotation and processing during the inference phase, and can choose whether to give the final classification result based on the confidence level. Only news headlines are used as evidence, and the Transformer decoder is used to combine self-attention and cross-attention to fuse news and evidence for prediction. Although data with labeled news categories is still required during the model training phase, other data of labeled evidence, such as user information, comment information, etc., is not required. This method can make the construction of the data set no longer require too much external information and features, and reduce the requirements for the dataset fields.

2. Multimodal News

In the field of multimodal news, more visual features are added. Better integration of features can maximize the domain coverage of labeled data and select more valuable data, thereby improving the generalization of fake news detection models in various fields. An unsupervised method is used to convert the multimodal data (text, communication network, etc.) of each article into a low-

dimensional vector to represent domain information in [19]. Build a heterogeneous network based on user homogeneity and domain-specific words to improve the representativeness of the selected samples in the field. One piece of news is selected from each group for manual annotation. The Locality Sensitive Hashing (LSH) method improves the F1 score of the fake news detection model by 24% under the premise of a fixed annotation budget.

Multimodal content can be extended to images, audio, video, etc., and more visual features can be added to integrate multimodal fake news features to make the data sampled by LSH more authoritative and representative. However, a method is proposed in [19] to balance data from different fields through multimodal data, laying a solid foundation for subsequent research.

3.3. Dynamic News Dataset Annotation

Regarding the construction and annotation of dynamic news datasets, considering the problem of insufficient annotated data in fake news detection, a framework that uses user feedback (such as user reports and comments) as weak supervision signals is proposed in [20]. Improving the performance of fake news detection through automatic labeling and reinforcement learning techniques. The model consists of three parts: an annotator, a Reinforced Data Selector (RDS), and a fake news detector. Firstly, the annotator extracts the features of user report feedback through the CNN model as the basis for annotating weak labels for news. And train the automatic annotator using some of the labeled data. The annotator is used to predict whether the unlabeled data is fake news. Secondly, RDS aims to filter out high-quality samples from weakly labeled data. This includes states, actions, and rewards. The detection model is trained using selected samples and original samples, respectively. After calculating the difference in accuracy between the two, if the selected sample helps improve the model effect, a positive reward is given, otherwise, a negative reward is given. Finally, the policy gradient method is used to update the sample selector so that it can capture samples of higher quality. The model performs better than machine learning and deep learning models, as shown in the training of the fake news detector in the third part. At the same time, dynamically adjusting the sample selection strategy to screen out high-quality samples can improve the generalization ability of the model.

4. Future Directions

(1) Real-time detection of fake news

Most current datasets lack the ability to capture dynamic changes in news, and current research on dynamic dataset annotation is also limited. Future research should integrate real-time crawler technology, reinforcement learning, and incremental learning to construct dynamically updated datasets so that detection models can adapt to the rapid evolution of news.

(2) Multimodal and cross-domain generalization issues

The changes in fake news are very complex, and a variety of disguised methods are emerging. In the future, it is necessary to integrate multimodal features further, explore the relationship between text and visual features, introduce user social network analysis, add knowledge graphs and external information, etc. However, when adding features to the model, the features must be well-screened and integrated. The features must be screened out through experiments to find the core parts, to maximize the performance of the model. At the same time, cross-domain datasets still need to be constructed to explore commonalities among multiple fields.

(3) Scarcity of multilingual and cross-cultural datasets

The significance of fake news detection is extraordinary for the whole world, but the current datasets are mainly in English. To address the scarcity of non-English datasets, it is necessary to build multilingual datasets. Different countries have different cultures and use words and sentences differently. When labeling and detecting, the model is prone to misjudgment due to the lack of this part of the data in its training data. At the same time, it is necessary to consider the impact of cultural background on the authenticity of news.

5. Conclusion

Starting from the background of fake news, the existing research on the construction of fake news datasets is reviewed in this paper. The existing literature is summarized and sorted from three perspectives: traditional manual annotation methods, semi-automatic annotation methods, and dynamic dataset annotation methods. Finally, areas for future improvement are proposed. Continuous optimization of dataset construction methods is essential to provide robust foundational support for fake news detection.

References

[1] Zhou Kaimin, Shu Chang, Li Binyang, et al. Early rumour detection. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 1614-1623.

[2] Hu Beizhe, Sheng Qiang, Cao Juan, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(20): 22105-22113.

[3] Mosallanezhad A, Karami M, Shu Kai, et al. Domain adaptive fake news detection via reinforcement learning. Proceedings of the ACM web conference 2022, 2022: 3632-3640.

[4] Zhang Litian, Zhang Xiaoming, Zhou Ziyi, et al. Reinforced adaptive knowledge learning for multimodal fake news detection. Proceedings of the AAAI conference on artificial intelligence, 2024, 38(15): 16777-16785.

[5] Al-Quayed F, Javed D, Jhanjhi N Z, et al. A Hybrid Transformer-Based Model for Optimizing Fake News Detection. IEEE Access, 2024, 12: 160822-160834.

[6] Qin Simeng, Zhang Mingli. Boosting generalization of fine-tuning BERT for fake news detection. Information Processing & Management, 2024, 61(4): 1-18.

[7] Mahmud T, Akter T, Aziz M T, et al. Integration of NLP and deep learning for automated fake news detection. 2024 Second International Conference on Inventive Computing and Informatics (ICICI), 2024: 398-404.

[8] Alghamdi J, Luo S, Lin Y. A comprehensive survey on machine learning approaches for fake news detection. Multimedia Tools and Applications, 2024, 83(17): 51009-51067.

[9] Kumar Y. Combating Misinformation: Insights into Datasets, Models and Evaluation Strategies for Fake News. 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON). IEEE, 2024: 1-4.

[10] Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic Detection of Fake News. Proceedings of the 27th International Conference on Computational Linguistics, 2018: 3391-3401.

[11] Zhang Amy X, Ranganathan A, Metz S E, et al. A structured response to misinformation: Defining and annotating credibility indicators in news articles. Companion Proceedings of The Web Conference 2018, 2018: 603-612.

[12] Bonet-Jover A, Sepúlveda-Torres R, Saquete E, et al. RUN-AS: a novel approach to annotate news reliability for disinformation detection. Language Resources and Evaluation, 2024, 58(2): 609-639.

[13] Raza S, Paulen-Patterson D, Chen Ding. Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data. Knowl Inf Syst 67, 2025: 3267-3292.

[14] Bonet-Jover A, Sepúlveda-Torres R, Saquete E, et al. Applying Human-in-the-Loop to construct a dataset for determining content reliability to combat fake news. Engineering applications of artificial intelligence, 2023, 126: 107152.

[15] Bonet-Jover A. Semi-automatic annotation proposal for increasing a fake news dataset in spanish. CEUR Workshop Proceedings, 2021.

[16] Paka W S, Bansal R, Kaushik A, et al. Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. Applied Soft Computing, 2021, 107: 1-13.

[17] Akhtar M M, Karunanayake I, Sharma B, et al. Towards Automatic Annotation and Detection of Fake News. 2023 IEEE 48th Conference on Local Computer Networks (LCN), 2023: 1-9.

[18] Yang Yuzhou, Zhou Yangming, Ying Qichao, et al. Search, Examine and Early-Termination: Fake News Detection with Annotation-Free Evidences. IOS Press, 2024: 1463-1470.

[19] Silva A, Luo L, Karunasekera S, et al. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(1): 557-565.

[20] Wang Yaqing, Yang Weifeng, Ma Fenglong, et al. Weak supervision for fake news detection via reinforcement learning. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(01): 516-523.