

A Multifaceted Investigation on Enhancing Fake News Detection

Xinnan Xue

Queen Mary School Hainan, Beijing University of Posts and Telecommunications, Hainan, China
20050115Xxn@bupt.edu.cn

Abstract. In the age of digitalization, information overload exacerbates the spread of misinformation, the rapid development of the Internet and social media has reconstructed the information communication pattern, and the proliferation of false information has posed a severe threat to social stability and public safety. Against this background, the demand for fake news detection has increased, and improving the accuracy and efficiency of rumor detection has become an important issue to be solved urgently. This paper systematically reviews the methods of strengthening fake news detection with the help of large language models (LLMs), covering data enhancement, knowledge-driven prompting, BERT-LSTM hybrid models, and hybrid models. Although these methods have improved the detection performance to varying degrees, they also face challenges such as insufficient model interpretability, limited applicability in real scenarios, and difficulty in coping with evolving false information strategies. Future research can focus on combining explainable artificial intelligence, strengthening domain adaptation, and promoting multimodal fusion and cross-lingual expansion to further optimize the effect of fake news detection.

Keywords: Large Language Models, False Information Detection, Knowledge-Driven Prompting, Data Augmentation.

1. Introduction

In the digital information era, the rapid development of the Internet and social media has reshaped the information dissemination ecosystem. These advancements have led to the proliferation of false information, posing significant challenges to social stability and public safety [1]. Therefore, improving the accuracy and efficiency of rumor detection has become a critical priority [2].

In the field of fake news detection, Large Language Models (LLMs) represent a cutting-edge technology with immense potential. Leveraging their natural language understanding capabilities, LLMs can rapidly parse massive volumes of text, demonstrating powerful strengths in natural language processing tasks and text generation. This has opened up entirely new pathways for fake news detection.

Significant progress has been made in enhancing fake news detection using LLMs. However, real-world fake news datasets often contain fewer fake news samples than genuine ones, leading to imbalanced distributions that bias classifier performance. Studies have shown that the "Rumor Large Language Model" (RumorLLM) fills the existing gap of LLMs specifically tailored for fake news detection. The paper also proposes a method combining RumorLLM and prompt engineering, which diversifies and enhances minority-class samples, thereby improving the model's ability to distinguish complex Artificial Intelligence (AI)-generated rumors [3].

Research has demonstrated the effectiveness of knowledge-driven prompting methods to enhance LLM performance in rumor detection tasks. The framework of Enhancing large language model capabilities for rumor detection with Knowledge-Powered Prompting (ELKP) integrates domain-specific terminology, relationships, and contextual information to guide LLMs in executing domain-specific tasks. Injecting domain knowledge enhances the model's capability and effectiveness in handling tasks within specific fields, making it more adaptive and scalable [4]. The method of Bidirectional Encoder Representations from Transformers (BERT) with Long Short-Term Memory (LSTM) networks to enhance the detection of misinformation using only textual content (BERT-LSTM), by combining the advantages of contextual understanding and sequential processing,

provides a more comprehensive technical solution for in-depth analysis of misinformation in mobile social networks [4].

The effective application of LLMs in financial news analysis has also provided practical cases for fake news detection research. For example, a fine-tuned Llama 2 LLM was used for multi-task financial news analysis, demonstrating its ability to perform structured analysis. Results showed that the fine-tuned Llama 2 model can deeply analyze texts and reveal complex styles and narratives [5].

This review paper aims to systematically summarize the methods of enhancing fake news detection using LLMs. The rest of this paper is structured as follows: 1) Section 2 describes the frameworks of different model approaches. 2) Section 3 provides discussion (challenges and future prospects) 3) Section 4 concludes the whole paper.

2. Method

2.1. Enhance Fake News Detection with Data Augmentation Methods

Lai proposed Rumor LLM, a data augmentation method for fake news detection, aiming to address the performance bias of classifiers caused by the imbalanced distribution of fake news samples in datasets. The preprocessing stage of the model involves using the Porter Stemming algorithm to extract stems from original news text for reducing lexical complexity, and applying the TF-IDF algorithm to calculate term frequency, word frequency in the current document, and inverse document frequency—thereby measuring the importance of words in the document collection. Keywords are then reconstructed based on the processed stems and TF-IDF values to form prompts and responses for model training.

The workflow is as follows: First, the original news text undergoes word segmentation and other processing. Then, the Porter Stemming and TF-IDF algorithms are used for preprocessing to obtain a keyword list. Next, prompts and responses are reconstructed based on this list, with the original rumor text serving as the response. The basic large language model is fine-tuned using methods like P-tuning V2 to obtain RumorLLM. Finally, RumorLLM combined with prompt engineering generates diversified small-category samples to achieve data augmentation. The overall framework consists of basic large language models (such as ChatGLM series, LLaMA series, etc.), Porter Stemming algorithm processing module, TF-IDF algorithm processing module, prompt and response reconstruction module, and P-tuning V2 fine-tuning module.

By combining RumorLLM with state-of-the-art classification models for data augmentation, this method effectively improves the category imbalance problem of datasets, significantly enhances F1 score and AUC-ROC, verifies the effectiveness of the new idea of "large language model + data augmentation", and strengthens the prediction capability of rumor detection models.

2.2. Enhance Large Language Model Capabilities for Rumor Detection with Knowledge-Driven Prompts

Yan et al. proposed a knowledge-driven prompt-enhanced LLM for rumor detection. The preprocessing stage of the model includes performing named entity recognition on the input text to extract key entities, and then constructing exploration prompts according to preset rules to retrieve and filter background information of key entities from the knowledge base, thereby enhancing prompt representation.

The workflow is: Key entities in the input text are first extracted through named entity recognition, followed by constructing exploration prompts, retrieving and filtering background knowledge of key entities from the knowledge base, inputting the enhanced prompts into the large language model for fine-tuning, and finally using a small neural network classifier to complete the classification task of rumor detection. The overall framework is composed of core modules such as exploration prompt construction, knowledge-prompt combination, large language model, and optimization objectives.

The innovations are mainly reflected in the following aspects: Constructing exploration prompts to strengthen the understanding of key entities in the text, obtain richer semantic representations, and

guide the large language model to carry out comprehensive text reasoning; proposing a knowledge-prompt combination module to dynamically learn external knowledge related to prompts and selectively insert it into prompt representations, enabling the model to focus more on semantically relevant content and improve performance; conducting multi-task joint optimization to achieve collaborative training of semantic representation extraction and discriminative judgment, enhancing the comprehensive performance of the model. This model effectively improves the accuracy of rumor detection and the adaptability of the model in data-scarce scenarios, providing new solutions for the field.

2.3. Tackling Misinformation in Mobile Social Networks A BERT-LSTM Approach for Enhancing Digital Literacy

Wang et al. proposed a BERT-LSTM large language model (LLM) for misinformation detection. The preprocessing stage of the model includes: performing word segmentation on original text data, splitting text into individual tokens, and introducing special tokens (such as [CLS]) to mark the beginning of the sequence; cleaning the text to remove illegal characters, links, irrelevant titles, etc.; for English content, removing punctuation and stop words; for Chinese content, performing word segmentation; merging news titles and contents to enable the model to consider complete contextual information during classification; and standardizing date information.

The workflow covers: The input layer converts preprocessed text data into vector representations, including the summation of word vectors, segment vectors, and position vectors. The BERT layer processes vectors through multiple Transformer layers, using self-attention mechanisms to capture bidirectional contextual information of the text and generate contextual embeddings. The LSTM layer processes the contextual embeddings generated by BERT to capture sequence dependencies in the text and outputs hidden state sequences. The output layer aggregates and classifies the hidden states of LSTM through fully connected layers and Softmax layers, outputting the probability distribution of the text being misinformation.

The overall framework consists of the following core modules: The BERT module based on the Transformer architecture, used to extract deep semantic features and contextual relationships of the text; the LSTM module, used to process sequence data and capture long-term dependencies in the text; and the classification module, composed of fully connected layers and Softmax layers, used for classification and prediction of text.

Innovations are mainly reflected in: Integrating the powerful contextual understanding ability of BERT with the sequence processing ability of LSTM to form a BERT-LSTM hybrid model, which can capture both semantic features and sequence dependencies of the text, effectively improving the accuracy of misinformation detection. The model achieved an accuracy of 93.51%, recall of 91.96%, and F1 score of 92.73% in experiments, showing significant advantages over traditional machine learning models and other deep learning models; enabling early detection of misinformation based solely on text content without relying on user participation data (such as likes, shares, etc.), solving the problem that traditional methods are difficult to detect in the early stage of misinformation propagation due to lack of user feedback data, and providing possibilities for timely intervention in misinformation; serving as an educational tool to improve digital literacy—user experiments show that the experimental group using the model has higher misinformation detection accuracy (89.4%), decision-making speed, and judgment confidence than the control group (74.2%), helping to cultivate users' critical thinking ability and providing new ideas for digital literacy education.

2.4. Enhance Disinformation Detection with Hybrid Models

Bohdan M. Pavlyshenko proposed a fine-tuned Llama 2 large language model (LLM) for false information and fake news detection. The preprocessing of the model involves collecting, integrating and dividing datasets, obtaining fake news datasets from the Kaggle repository, supplementing them with Russian propaganda narrative texts from Vox Ukraine, constructing a comprehensive dataset for fine-tuning, and dividing it into training and validation sets at a ratio of 4:1 to ensure the effectiveness

of model training and validation; designing prompt words, which are generated by experts or with the help of other LLMs such as ChatGPT and GPT-4, to clarify the task direction of the model, such as "analyzing the text from the perspective of exposing false information narratives", so as to provide clear task guidance for model fine-tuning; optimizing the quantization model, adopting 4-bit quantization technology to reduce the video memory occupation of the model and adapt to low-cost GPU resources.

The workflow includes model selection and initialization, selecting the basic model as Llama 2 (specifically meta-llama/Llama-2-7b-chat-hf), initializing the model based on the PEFT (parameter efficient fine-tuning) framework, and loading 4-bit quantization parameters to optimize resource consumption; fine-tuning training, using the SFTTrainer (supervised fine-tuning trainer) in the trl library for model training, setting key parameters (learning rate $5e-4$, training rounds 10, maximum sequence length 2048, gradient accumulation steps 2, etc.), and only fine-tuning part of the model parameters through LoRA (low-rank adaptation) technology to reduce the calculation cost while retaining the core capabilities of the model; multi-task test verification, using the fine-tuned model for a number of task tests, including false information and propaganda narrative recognition, fact checking, fake news detection, manipulative text analysis, named entity and its emotion extraction, and obtaining the model analysis results by inputting prompt words with a specific structure including system role definition and task instructions; result evaluation and optimization, qualitatively evaluating the structured text and JSON format results output by the model, and proposing to optimize the model performance by expanding the training dataset screened by experts and adopting RLHF (reinforcement learning based on human feedback) for the inaccurate problems in the analysis.

The overall framework includes a data preprocessing module, which includes Kaggle fake news datasets, Vox Ukraine propaganda narrative texts, and the divided training and validation sets, providing training and validation data support for the model. At the same time, the integration and division of the datasets, the design of prompt words and the 4-bit quantization of the model ensure that the data and model are adapted to the subsequent fine-tuning process; a fine-tuning module, with PEFT/LoRA technology as the core, combined with the SFTTrainer trainer, realizes the parameter efficient fine-tuning of the Llama 2 model, only updating a small number of low-rank matrix parameters to reduce the calculation cost; a task execution module, which receives the text and prompt words input by users, calls the fine-tuned Llama 2 model to complete the multi-task analysis such as false information detection and emotional entity extraction, and outputs structured text or JSON format results; an evaluation and optimization module, which qualitatively evaluates the output results of the model, gives feedback on problems and guides the subsequent improvement operations such as expanding the training dataset and optimizing RLHF.

The innovations are mainly reflected in the adoption of the PEFT/LoRA framework, which only fine-tunes a small number of low-rank matrix parameters of the Llama 2 model instead of all parameters, and combined with 4-bit quantization technology, enables the LLM with 7 billion parameters to complete fine-tuning on low-cost GPUs (such as Google Colab), greatly reducing the calculation and storage costs, while the performance is close to that of the full fine-tuning model; the model supports multiple tasks such as false information recognition, fact checking, and named entity emotion extraction at the same time, and the output results include structured text and JSON format data, expanding the application scenarios of the model in subsequent prediction tasks. The model effectively improves the multi-task adaptability and resource utilization efficiency of fake news detection, and its structured output provides a data foundation for the subsequent construction of more complex false information analysis systems, and provides a new solution idea of "low-cost fine-tuning + multi-task integration + scenario adaptation" for the fake news detection field.

3. Discussion

3.1. Challenges

Insufficient Interpretability of Detection Models: The lack of interpretability of advanced models (especially hybrid architectures and large language models) is a key challenge in the field of false information detection. As shown in [4], the BERT-LSTM hybrid model integrates the contextual understanding ability of BERT and the sequence processing advantage of LSTM, and although it achieves a high accuracy of 93.51%, its decision-making process remains in a "black box" state. Traditional machine learning models can trace prediction bases through feature weights, while deep learning models such as BERT-LSTM and GPT-3 cannot clearly present key decision-making logic. For example, [5] mentions that even if the fine-tuned Llama 2 model can generate false information analysis results, it is difficult to determine whether the classification results are dominated by the emotional tendencies of named entities, semantic patterns, or other language clues. This insufficient interpretability seriously affects trust. In public health scenarios, if the model determines that a piece of vaccine-related content is false information, the regulatory agency cannot clearly explain the basis, making it difficult to persuade the public to accept the conclusion. In addition, the ELKP framework proposed in [4], although it performs better than the baseline model on the Twitter15/16 dataset by combining knowledge-driven prompts with LLaMA, the criteria for its dynamic knowledge injection module to screen "relevant" external knowledge lack transparency, and it is impossible to verify whether the model relies on accurate and unbiased information for decision-making.

Limited Applicability of Models in Real Scenarios: The applicability of existing models in the real scenario of dynamic mobile social networks is insufficient, mainly reflected in three aspects. First, the problem of data imbalance is prominent. [3] points out that in real-world fake news datasets, the number of fake news samples is usually 1/3 to 1/4 of that of real news, and traditional data augmentation cannot capture the stylistic features of rumors. Although generative methods such as RumorLLM are used to balance the dataset, the generated content is difficult to fully simulate the diversity of real false information. Second, the ability to deal with multimodal false information is missing. Most current models, such as BERT-LSTM and fine-tuned Llama 2, only focus on text and ignore visual clues such as images and videos. In real scenarios, false information often spreads in the form of "text + tampered images". [3] admits that RumorLLM cannot process audio-visual false information, limiting its application on platforms such as TikTok and Instagram. Third, cross-lingual and cross-cultural adaptability is poor. Models are mainly tested on English and Chinese texts, while false information is global. For example, for Russian false information targeting Ukraine, if the model is only trained on English datasets, it will fail to detect due to insufficient training data and cultural differences when facing low-resource languages such as Ukrainian and Bulgarian.

Difficulty in Coping with Evolving False Information Strategies: False information creators continue to adjust their strategies to avoid detection, bringing huge challenges to static models. [4] points out that rumors increasingly use "unconventional semantic structures" such as irony and veiled hints, and disguise legitimacy by borrowing hot topics (such as epidemics and elections). For example, during the 2024 U.S. presidential election, false information may integrate real political events with forged remarks, making it difficult for models trained on 2016 election data to identify. In addition, AI-generated false information increases complexity. RumorLLM in [appls-ci-14-03532.pdf](#) can generate rumor-style texts, meaning that bad actors may use similar tools to create more complex false information and break through the detection threshold of existing models. It also mentions that if the training data of the fine-tuned Llama 2 model does not cover subtle political propaganda content (such as the controversial remarks of Hungarian officials on Ukraine), misclassification may occur, highlighting the vulnerability of the model to emerging false information strategies.

3.2. Future Prospects

Combining Explainable AI to Improve Model Transparency: To solve the "black box" problem of models, explainable AI technology needs to be integrated into false information detection [6-8]. For

hybrid architectures such as BERT-LSTM, attention visualization can be used to mark the text fragments (such as specific sentences and keywords) that the model prioritizes when classifying. For example, a heat map can clarify that the model determines a piece of COVID-19 false information because of the unfounded statement of "vaccine side effects" rather than the neutral description of symptoms. For large language models, "thinking chain" prompts can be designed to require the model to output the basis for classification, such as "This content is determined to be false because it contradicts the WHO mask-wearing guidelines". In addition, the named entity emotion extraction method can be expanded into an "explainable report" to associate model predictions with specific entities, thereby improving the credibility and availability of the model in education, public policy, and other fields.

Focusing on Domain Adaptation and Strengthening Performance in Specific Scenarios [9, 10]: Future models need to prioritize domain adaptation to improve performance in high-value scenarios. In the field of public health, models can be trained based on domain datasets such as epidemic false information and vaccine rumors, integrating professional knowledge bases such as WHO guidelines and PubMed studies, and at the same time transforming relevant frameworks into the logic of "first verifying whether it conforms to CDC recommendations, then classifying" to improve detection accuracy. In the field of electoral politics, for false information in the electoral cycle, the model can be fine-tuned based on real-time data from social platforms, incorporating background information such as candidates, political parties, and electoral rules to distinguish between reasonable criticism and false accusations (such as "The content of 'election fraud' is false because there is no evidence from the state election commission"); generative models can also be used to generate synthetic rumors in specific fields (such as false information related to campaign funds) to help the model adapt to time-sensitive false information in the electoral cycle.

Promoting Multimodal Fusion and Cross-Lingual Expansion: To meet the needs of real scenarios, models need to achieve multimodal fusion and cross-lingual detection capabilities. In terms of multimodality, generative models can be expanded to combine text generation and visual synthesis (such as generating tampered images and supporting false text) to train multimodal classifiers, and cross-modal attention mechanisms can be used to associate text expressions with visual clues, such as verifying whether the "protest scene" in news pictures is consistent with the location and time mentioned in the text. In terms of cross-lingual, "translation enhancement" strategies (such as translating low-resource language false information into common languages for pre-training, and then fine-tuning with local data) can be adopted for models such as BERT-LSTM and Llama 2, combined with multilingual knowledge bases (such as multilingual Wikipedia data). Existing fine-tuned models can already analyze false information in specific languages, indicating that cross-lingual transfer learning is feasible. After further optimization, it can cover low-resource language areas such as Southeast Asia and Africa to achieve global false information

4. Conclusion

This paper reviews the methods of fake news detection based on large language models, covering strategies such as data augmentation, knowledge-driven prompting, BERT-LSTM hybrid models, and hybrid models. Practices have shown that although these schemes improve detection efficiency, with the rapid development of information dissemination, there are limitations in model interpretability, real-world scenario adaptation, and dynamic rumor response. Future research needs to advance from the dimensions of explainable AI integration, domain adaptability enhancement, and multimodal cross-lingual detection expansion to construct a more complete rumor detection system.

References

- [1] Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. Combating fake news: A survey on identification and mitigation techniques. *ACM transactions on intelligent systems and technology (TIST)*. 2019 Apr 12; 10 (3): 1-42.

- [2] Zhou X, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*. 2020 Sep 28; 53 (5): 1-40.
- [3] Lai J, Yang X, Luo W, Zhou L, Li L, Wang Y, Shi X. Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach. *Applied Sciences*. 2024 Apr 22; 14 (8): 3532.
- [4] Wang J, Wang X, Yu A. Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy. *Scientific Reports*. 2025 Jan 7; 15 (1): 1118.
- [5] Pavlyshenko BM. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv: 2309.04704*. 2023 Sep 9.
- [6] Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing 2019 Sep 30* (pp. 563-574). Cham: Springer International Publishing.
- [7] Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, Ranjan R. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys*. 2023 Jan 13; 55 (9): 1-33.
- [8] Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. Explainable AI methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers 2020 Jul 18* (pp. 13-38). Cham: Springer International Publishing.
- [9] Farahani A, Voghoei S, Rasheed K, Arabnia HR. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*. 2021 Oct 30: 877-94.
- [10] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*. 2006; 19.