Engineering Applications of Artificial Intelligence 163 (2026) 112959

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A text-based hybrid transfer learning model for ternary classification in health misinformation detection

Jia Luo [a,b,c], Yang Yang [b,*], Xiaoye Feng [b], Didier El Baz [d,e]

[a] Interdisciplinary Faculty of Science and Engineering, Shimane University, Shimane, 690-8504, Japan
[b] College of Economics and Management, Beijing University of Technology, Beijing, 100124, China
[c] Chongqing Research Institute, Beijing University of Technology, Chongqing, 401121, China
[d] LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, 31400, France
[e] College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, China

ARTICLE INFO

Keywords:
Health misinformation detection
Ternary classification problem
Transfer learning
hybrid model

ABSTRACT

Detecting health misinformation is essential for protecting public health and ensuring effective communication during health crises. Significant attention has been devoted to health misinformation detection following the Coronavirus Disease 2019 (COVID-19) pandemic. Various approaches have been developed to automatically address health misinformation, often framing the problem as a binary classification task. However, these methods tend to overlook the complexity and fluidity of health-related information, where ongoing scientific research or incomplete data can complicate the definitive classification of certain claims. This paper approaches health misinformation detection as a ternary classification problem, categorizing content as uncertain, false, or true. A hybrid transfer learning model is proposed to effectively detect health misinformation by leveraging the linguistic features of general misinformation and combining multimodal features with an attention mechanism. The model is trained on both Chinese and English datasets, resulting in accuracy improvements of 6.75 % and 3.4 %, respectively.

1. Introduction

Health misinformation refers to false or misleading information related to health, often shared either intentionally or due to misunderstanding (Chou et al., 2018) (Swire-Thompson and Lazer, 2020). Its impact has become more pronounced with the rise of social media platforms and online communication, which enable the rapid spread of unverified content to a global audience. Unlike misinformation in other domains, health misinformation poses significant risks, including undermining trust in healthcare systems, delaying medical treatments, and promoting harmful practices. Its prevalence has been particularly evident during global health emergencies, such as the COVID-19 pandemic, where false claims about cures, vaccines, and prevention methods spread quickly, often fueled by fear, uncertainty, and a lack of reliable information (Cinelli et al., 2020). This misinformation complicates public health efforts, leading to vaccine hesitancy, non-adherence to safety measures, and increased public anxiety.

Detecting health misinformation is critical to safeguarding public health and ensuring effective communication during health crises. Timely and accurate identification of false or misleading health information can mitigate its harmful effects, such as the spread of incorrect medical advice, erosion of trust in science, and the perpetuation of unsafe behaviors. Effective detection not only supports public health campaigns but also counters the influence of misinformation and encourages informed decision-making among the public. However, detecting health misinformation presents significant challenges (Friggeri et al., 2014) (Pennycook and Rand, 2018) (Vosoughi et al., 2018). Firstly, the sheer volume of content generated daily on digital platforms makes manual verification impractical. Secondly, health misinformation often spreads rapidly due to its sensationalist or emotionally charged nature, which increases its likelihood of being shared or clicked on. Finally, health-related misinformation frequently involves complex medical terminology, nuanced concepts, and rapidly changing scientific evidence, making it difficult for automated systems to accurately identify false claims without domain-specific expertise.

Manual fact-checking remains a cornerstone of health

This article is part of a special issue entitled: AI for HPC systems - MGE: Grégoire G. Danoy published in Engineering Applications of Artificial Intelligence.

* Corresponding author.
E-mail addresses: jialuo@riko.shimane-u.ac.jp (J. Luo), yangy@bjut.edu.cn (Y. Yang), fengxiaoye2002@163.com (X. Feng), elbaz@laas.fr (D. El Baz).

https://doi.org/10.1016/j.engappai.2025.112959
Received 31 December 2024; Received in revised form 2 October 2025; Accepted 23 October 2025

misinformation detection, valued for its high accuracy. Organizations like FactCheck.org (FactCheck.org) and Health Feedback (Health Feedback) employ experts to evaluate the veracity of health claims. While this approach yields reliable results, it is labor-intensive and struggles to keep pace with the rapid spread of misinformation on social media platforms. In contrast, automated techniques and hybrid approaches have gained significant attention as scalable alternatives to manual methods. Automated techniques leverage advancements in natural language processing and machine learning to identify misinformation efficiently (Zhou and Zafarani, 2020). Hybrid approaches, which combine automated techniques with expert validation, further highlight the role of automation in scaling up misinformation detection (Nguyen and Shirai, 2015). However, the accuracy of automated techniques is still lower, and detecting health misinformation becomes even more challenging than general misinformation due to the complexity of health-related topics, the evolving nature of medical knowledge, and the use of persuasive or emotionally charged language in health-related content (Di Sotto and Viviani, 2022).

Several approaches have been employed to automatically combat general misinformation, with particular attention given to health misinformation detection following the COVID-19 pandemic (Papanikou et al., 2024). These methods typically approach the problem as a binary classification task, where the goal is to distinguish between true and false information (Schlicht et al., 2024). While this approach has been widely adopted, it overlooks the inherent challenges involved in making definitive judgments about the veracity of information, especially in a global health crisis. To address this, researchers emphasize the importance of framing health misinformation detection as a ternary classification problem, categorizing content as uncertain, false, or true (Luo et al., 2023). Uncertain refers to information that general healthcare workers cannot judge independently, while false refers to information they can identify as false, and true refers to information they can identify as true. This framework acknowledges the fluidity and complexity of health-related information, where ongoing scientific investigations or incomplete data can make it difficult to classify certain claims definitively (Shahi et al., 2021). Moreover, Incorporating an uncertain category enables more nuanced evaluations and reduces the risk of overgeneralization, which can increase the accuracy and reliability of automated techniques.

Therefore, this paper is driven by the need for ternary classification in health misinformation detection. A hybrid transfer learning model is developed to effectively categorize health misinformation as uncertain, false, or true, achieving a higher accuracy rate by fully leveraging general misinformation linguistics features.

The main contributions of our work are summarized as follows:

1. A pre-trained model is utilized to classify general misinformation as false or true, incorporating health-related keywords into the dataset.
2. A fine-tuned model is developed by combining BERT, TextCNN, and fastText with an attention mechanism for ternary classification, categorizing health misinformation as uncertain, false, or true.
3. The hybrid transfer learning model is evaluated on both Chinese and English datasets, achieving accuracy improvements of 6.75 % and 3.4 %, respectively.

The subsequent sections of this paper are structured as follows. Section 2 provides an overview of relevant literature in the field. Section 3 introduces a hybrid transfer learning model for ternary classification in health misinformation detection. Section 4 presents the experimental results along with statistical analysis. Finally, Section 5 offers the conclusions.

2. Related works

Misinformation on the Internet is generally treated as fake news or rumors (Bondielli and Marcelloni, 2019). Many natural language processing and machine learning based methods have been proposed to detect misinformation (Alghamdi et al., 2024a) while deep learning-based models have attracted significant attention recently. Arunthavachelvan et al. (2024) introduced a deep neural network approach integrating linguistic and psychological features to evaluate the truthfulness of news articles. Chen et al. (2024a) proposed a deep semantic-aware graph convolutional network designed for Cantonese rumour detection by integrating global and local semantic features. Chen et al. (2024b) developed a syntactic multi-level interaction network incorporating syntactic dependency relationships to enhance rumour identification. Ozcelik et al. (2025) presented a misinformation detection framework by leveraging user communities, transformer-based models, and contrastive learning. Alghamdi et al. (2024b) utilized fine-tuned BERT and cross-level cross-modality attention networks to enhance fake news detection. Praseed et al. (2023) addressed an ensemble method combining three fine-tuned pre-trained transformer models to increase the efficiency of fake news detection in Hindi. Alghamdi et al. (2024c) outlined a neural-based framework using enhanced hierarchical convolutional attention networks to detect both fake news content and spreaders. Raja et al. (2024) presented a fake news detection model for Dravidian languages, leveraging contextualized word embeddings and a hybrid multiscale residual CNN-BiLSTM architecture.

On one hand, significant attention has been devoted to health misinformation detection following the COVID-19 pandemic, with many automated techniques developed to categorize health misinformation during the crisis as either true or false. Xia et al. (2023) outlined a hybrid CNN-BiLSTM-AM model with an outlier knowledge management framework for detecting COVID-19 fake news. Alghamdi et al. (2023) evaluated transformer-based models for COVID-19 fake news detection where CT-BERT + BiGRU outperformed others. Chen et al. (2023) explored various deep learning model frameworks for detecting COVID-19 fake news, with BiLSTM yielding the best performance. Hussain et al. (2025) enhanced COVID-19 misinformation detection in Urdu by a cascaded group multi-head attention model. Hajek et al. (2024) proposed a news representation model that integrates information-seeking behavior and linguistic features to enhance fake news detection during the COVID-19 pandemic. On the other hand, health misinformation detection outside the COVID-19 context has also seen notable advancements. Sicilia et al. (2021) addressed micro-level rumour detection in health-related posts by introducing a rule-based space characterization filter. Zhao et al. (2021) presented a health misinformation detection model that incorporates both central-level and peripheral-level features to improve detection accuracy. Liu et al. (2019) designed a machine learning-based framework for detecting health-related misinformation by analyzing reliable and unreliable health information.

To our knowledge, the majority of automated techniques for detecting health misinformation have framed it as a binary classification problem, given that most datasets in this field are labeled as true or false (Murayama, 2021). Nonetheless, some studies have divided records into 3–5 categories to better understand health misinformation, especially after the COVID-19 pandemic. Sicilia et al. (2018) proposed a Zikavirus-related rumour detection dataset where tweets were manually labeled as rumour, non-rumour, or unknown. Shang et al. (2024) restructured a dataset from the cancer domain by mapping the numerical credibility ratings to categorical labels: 0–1 as false, 2–3 as mixed, and 4–5 as true. Haouari et al. (2021) presented an Arabic COVID-19 Twitter dataset, with each tweet labeled as true, false, or other. Cheng et al. (2021) compiled an English COVID-19 rumour dataset from news and tweets, which were manually classified as true, false, or unverified. Luo et al. (2021) collected Chinese infodemic content from Weibo and WeChat during COVID-19, classifying entries as true, false, or questionable after adjustments. Kim et al. (2021) created a dataset of English claims and related tweets, categorizing them into four types: COVID true, COVID fake, non-COVID true, and non-COVID fake. Dharawat et al.

(2022) released a dataset to assess health risks in COVID-19-related social media posts, categorizing English tweets into five groups: real news/claims, not severe, possibly severe, highly severe, or refutes/rebuts misinformation.

Numerous automated techniques, particularly deep learning-based methods, have been proposed for misinformation detection and have garnered significant attention. With the growing public focus on health misinformation following the COVID-19 pandemic, models specifically designed for health misinformation detection have been well developed. However, most of these models are limited to classifying health misinformation as either true or false. Given the complexity of health misinformation, it is essential to treat it as a multiclass classification problem. The aforementioned datasets, which categorize health misinformation into 3–5 groups, provide a foundation for research in this field. In this context, this paper focuses on developing a hybrid transfer learning model for ternary classification in health misinformation detection to achieve better accuracy than conventional models.

3. Methodology

3.1. Hybrid architecture

The hybrid architecture is proposed as the amount of health misinformation data is relatively limited, making it difficult to train a highly accurate model from scratch. By leveraging a pre-trained model on general misinformation and health-related keywords, the model can effectively transfer knowledge and capture health-specific patterns. Combining simpler model structures in a the fine-tuned stage allows the model to obtain multimodal features while avoiding overfitting, which can occur in more complex models when the available data is insufficient.

The overall design of the hybrid model is illustrated in Fig. 1. The pre-trained model is used for binary classification, categorizing misinformation as either "false" or "true". In this stage, health misinformation is preprocessed using TF-IDF to extract important keywords, and BERT embeddings are generated for both health-related keywords and general misinformation. The key component connecting the pre-trained model to the fine-tuned model is the output produced by the BERT layer from the general misinformation text, which is fused with the BERT embeddings for both health-related keywords and general misinformation. In the fine-tuned model, the focus shifts exclusively to health misinformation records, introducing an "uncertain" category for records that cannot be clearly classified as "false" or "true" by healthcare workers. This stage fine-tunes the model to classify health misinformation with greater precision, using BERT, TextCNN, and fastText embeddings. These heterogeneous features are then fused through an attention mechanism. Finally, the neural network head outputs one of three labels: "uncertain", "false", or "true".

3.2. Pre-trained model

3.2.1. Keyword extraction from health misinformation records

In the initial stage, the health misinformation records (denoted as $D = \{d_1, d_2, \cdots, d_n\}$, where each d_i is a health misinformation record) are processed using the TF-IDF (Robertson, 2004) method to extract the top 10 % of the most important keywords. Each health misinformation record d_i is tokenized into a list of terms using Jieba (https://github.com/jsrpy/Chinese-NLP-Jieba) for Chinese or regular expressions (Bird and Klein, 2006) for English text. The TF-IDF algorithm is applied to the tokenized health misinformation records where the TF-IDF score for each term t in document d_i is calculated as:

$$\mathrm{TF-IDF}(t, d_i, D) = \mathrm{TF}(t, d_i) \times \mathrm{IDF}(t, D) \tag{1}$$

where $\mathrm{TF}(t, d_i) = \frac{\text{Count of term } t \text{ in } d_i}{\text{Total number of terms in } d_i}$, $\mathrm{IDF}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$ with N denoting the total number of health misinformation records in D and $|\{d \in D : t \in d\}|$ representing the number of records containing the term t.

TF-IDF is a statistical method that measures the importance of a term within a document relative to the entire corpus. Terms that appear frequently in one document but rarely across the entire corpus are assigned higher weights. TF-IDF is adopted in the hybrid model as it is simple, interpretable, and effective for extracting representative keywords in health misinformation records, which provides a strong foundation for subsequent embedding and classification tasks.

After calculating the TF-IDF scores for each term across all records in D, the top 10 % of terms with the highest TF-IDF values are selected and considered as health misinformation keywords that provide the most relevant information about the misinformation content.

3.2.2. Embeddings for health misinformation keywords and general misinformation

Each keyword t_i from the set of health misinformation keywords (denoted as $X_{\text{Healthwords}} = \{t_1, t_2, \cdots, t_k\}$) is passed through the BERT model (Devlin et al., 2019) to produce a corresponding contextual embedding e_i. The BERT embeddings for health misinformation keywords are calculated as:

$$E_{\text{BERT-Healthwords}} = \mathrm{BERT}_{\text{embedding}}(X_{\text{Healthwords}}) = [e_1, e_2, \cdots, e_k] \in \mathbb{R}^{k \times H} \tag{2}$$

where e_i is the embedding vector for the i-th keyword t_i, $X_{\text{Healthwords}}$ represents the input corresponding to the selected health misinformation keywords, k denotes the number of selected keywords, and H is the hidden dimension of the BERT embedding.

The general misinformation text $X_{\text{Gen},j}$ is a sequence of tokens for the j-th instance. The BERT model processes $X_{\text{Gen},j}$ to generate a single contextual embedding for the entire sequence. The BERT embedding for the general misinformation text is calculated as:

$$E_{\text{BERT-Gen},j} = \mathrm{BERT}_{\text{embedding}}\left(X_{\text{Gen},j}\right) \in \mathbb{R}^{L_j \times H} \tag{3}$$

where $X_{\text{Gen},j}$ represents the sequence of general misinformation text for the j-th instance and L_j is its sequence length.

BERT is a pre-trained deep learning model that captures bidirectional contextual relationships between words. Unlike traditional word embeddings, BERT dynamically adjusts word representations based on surrounding context, making it highly effective for detecting subtle linguistic cues. BERT is selected in the hybrid model as health misinformation often relies on nuanced wording, and BERT provides richer contextual embeddings that can capture such semantics.

The embeddings for general misinformation text and health misinformation keywords are further fused to create a unified representation. The fusion is performed by concatenating the embedding of the general misinformation text $E_{\text{BERT-Gen},j}$ with the embeddings of each health misinformation keyword e_i, and the result is flattened into a single vector. The fused feature vector for the j-th general misinformation record is calculated as:

$$F_j = \mathrm{Flatten}\left(\cup_{i=1}^{k}\left(E_{\text{BERT-Gen},j} \text{ concat } e_i\right)\right) \in \mathbb{R}^{\left(L_j + k\right) \times H} \tag{4}$$

This fusion provides a comprehensive representation by integrating the general misinformation context with the specific health misinformation keywords, offering the model richer information for subsequent fine-tuning and classification tasks.

3.2.3. Binary classification using the pre-trained model

The key component connecting the pre-trained model to the fine-tuned model is the output produced by the BERT layer from the general misinformation text. It consists of contextualized token embeddings and the pooled output. The pooled output is typically obtained by extracting the [CLS] token output from BERT, which is applied to the general misinformation text, and is calculated as:

$$\text{pooled\_output}_j = \text{BERT}_{\text{pooled}}(X_{\text{Gen},j}) \in \mathbb{R}^H \tag{5}$$

The model then fuses the pooled output from the general misinformation text with the fused feature vector F_j. The combined feature vector is finally passed through a fully connected layer for binary classification. The prediction is calculated as:

$$\widehat{y}_{\text{binary}} = \sigma\Big(fc\Big(\Big[\text{pooled\_output}_j, F_j\Big]\Big)\Big) \tag{6}$$

where fc is the fully connected layer, σ is the sigmoid activation function, and $\widehat{y}_{\text{binary}}$ is the predicted probability for the "true" class in the binary classification task.

3.3. Fine-tuned model

3.3.1. Feature extraction

The output produced by the BERT layer in the pre-trained model from the general misinformation text is transferred to the fine-tuned model. The contextualized token embeddings and the pooled output for a batch of health misinformation text are calculated as:

$$\text{pooled\_output} = \text{BERT}_{\text{pooled}}(X_{\text{Health}}) = P \in \mathbb{R}^{B\times H} \tag{7}$$

$$\text{contextualized\_token\_embedding} = \text{BERT}_{\text{encoded}}(X_{\text{Health}}) = E \in \mathbb{R}^{B\times L\times H} \tag{8}$$

where X_{Health} represents the sequence of tokens for the input batch of health misinformation instances, B is the batch size, L is the sequence length.

The output from the pooled output is then passed through a dropout layer while the output of the contextualized token embeddings is further processed by the TextCNN (Kim, 2014) pooling layer and the fastText (Joulin et al., 2017) pooling layer, calculated as:

$$P_{\text{drop}} = \text{Dropout}\ (P) \in \mathbb{R}^{B\times H} \tag{9}$$

$$C_1 = \text{AdaptiveMaxPool1D}(\text{Conv1D}(E, k_1)) \in \mathbb{R}^{B\times d_c} \tag{10}$$

$$C_2 = \text{AdaptiveMaxPool1D}(\text{Conv1D}(E, k_2)) \in \mathbb{R}^{B\times d_c} \tag{11}$$

$$F_{\text{mean}} = \text{AdaptiveAvgPool1D}(E) \in \mathbb{R}^{B\times H} \tag{12}$$

$$F_{\text{max}} = \text{AdaptiveMaxPool1D}(E) \in \mathbb{R}^{B\times H} \tag{13}$$

where P_{drop} represents the output after applying dropout to regularize the pooled output. C_1 a nd C_2 are the outputs of the TextCNN pooling layer, which first applies convolutional operations with kernel sizes k_1 and k_2 to the token embeddings and then performs adaptive max pooling. d_c denotes the number of convolutional filters in each TextCNN layer. F_{mean} and F_{max} are the outputs of the fastText pooling layer, which applies adaptive average pooling and adaptive max pooling to the token embeddings.

TextCNN is a convolutional neural network designed for text classification. It applies convolutional filters with different kernel sizes to capture local n-gram features and then applies pooling to select the most salient signals. This makes TextCNN highly effective at detecting phrase-level and local semantic patterns in misinformation. FastText enriches word vectors with subword information and applies pooling mechanisms to capture overall sequence-level features. It is efficient and performs well in limited-data scenarios, making it suitable for health misinformation tasks where training data is relatively scarce. TextCNN and fastText are choosen in the hybrid model as they complement BERT. While BERT provides deep contextual embeddings, TextCNN excels at extracting localized features, and fastText captures broader sequence-level information. This combination enables a multi-level representation that balances complexity and efficiency.

3.3.2. Ternary classification using the fine-tuned model

The final feature representation is obtained by stacking all processed outputs into a sequence and fusing them via an attention mechanism, calculated as:

$$F = \left[P_{\text{drop}}, \widetilde{C}_1, \widetilde{C}_2, F_{\text{mean}}, F_{\text{max}}\right] \in \mathbb{R}^{B\times 5\times H} \tag{14}$$

$$F_{\text{att}} = \sum_{i=1}^{5} \alpha_i F_i \in \mathbb{R}^{B\times H} \tag{15}$$

$$\alpha_i = \frac{\exp\ (\text{score}(F_i))}{\sum_{j=1}^{5} \exp\ (\text{score}(F_j))} \tag{16}$$

where F denotes the set of feature vectors obtained from the dropout, TextCNN, and fastText pooling layers. $\widetilde{C}_1, \widetilde{C}_2$ are TextCNN features projected into the hidden space of dimension H. α_i represents the attention weight assigned to the i-th feature vector. $\text{score}(F_i) = v^\top \tanh\ (WF_i)$ is the attention scoring function with learnable parameters $W \in \mathbb{R}^{H\times H}$ and $v \in \mathbb{R}^H$.

Attention is a mechanism that dynamically assigns weights to different input features based on their relevance to the task. It enables the model to selectively emphasize informative components while downplaying less relevant ones. Attention is incorporated in the hybrid model as it enhances feature fusion by adaptively highlighting key signals across BERT, TextCNN, and fastText features, improving accuracy and robustness in ternary classification tasks.

Finally, the fused feature F_{att} is passed through a fully connected layer for ternary classification. The prediction is calculated as:

$$\widehat{y}_{\text{ternary}} = \phi(fc(F_{\text{att}})) \tag{17}$$

where ϕ is the softmax activation function and $\widehat{y}_{\text{ternary}}$ is the predicted probabilities for all classes in the ternary classification task.

The combination of these feature extraction and fusion steps effectively captures multiple levels of information, including contextual, local, and global features. This comprehensive approach strengthens the model's ability to perform ternary classification in health misinformation detection, resulting in improved accuracy and robustness.

4. Experiments

4.1. Setup

4.1.1. Experimental environment

The experimental environment was developed in Python and executed on a server equipped with an NVIDIA GeForce RTX 4090, featuring 24 GB of GDDR6X memory. The deep learning model was built using Anaconda, a Python distribution for scientific computing, while PyTorch, a widely-used open-source machine learning framework, was employed to design and implement the model architecture.

4.1.2. Datasets

Concerning health misinformation records, two balanced datasets proposed in (Luo et al., 2024) are utilized. These datasets consist of social media textual records that have been refined with annotations from healthcare workers and categorized into three groups, with labels 0, 1, and 2 corresponding to uncertain, false, and true, respectively. The Chinese dataset comprises 1055 records, with 435 labeled as uncertain, 281 as false, and 339 as true, collected from manually verified Weibo posts, the WeChat mini-program "Jiaozhen", and other authoritative sources. The English dataset contains 2490 records, evenly distributed across the three categories, sourced from public fact-verification websites, the Twitter API, and other social media platforms. 15 representative examples with their labels are shown for the Chinese dataset in Fig. 2 and for the English dataset in Fig. 3. For model tuning, 10 % of

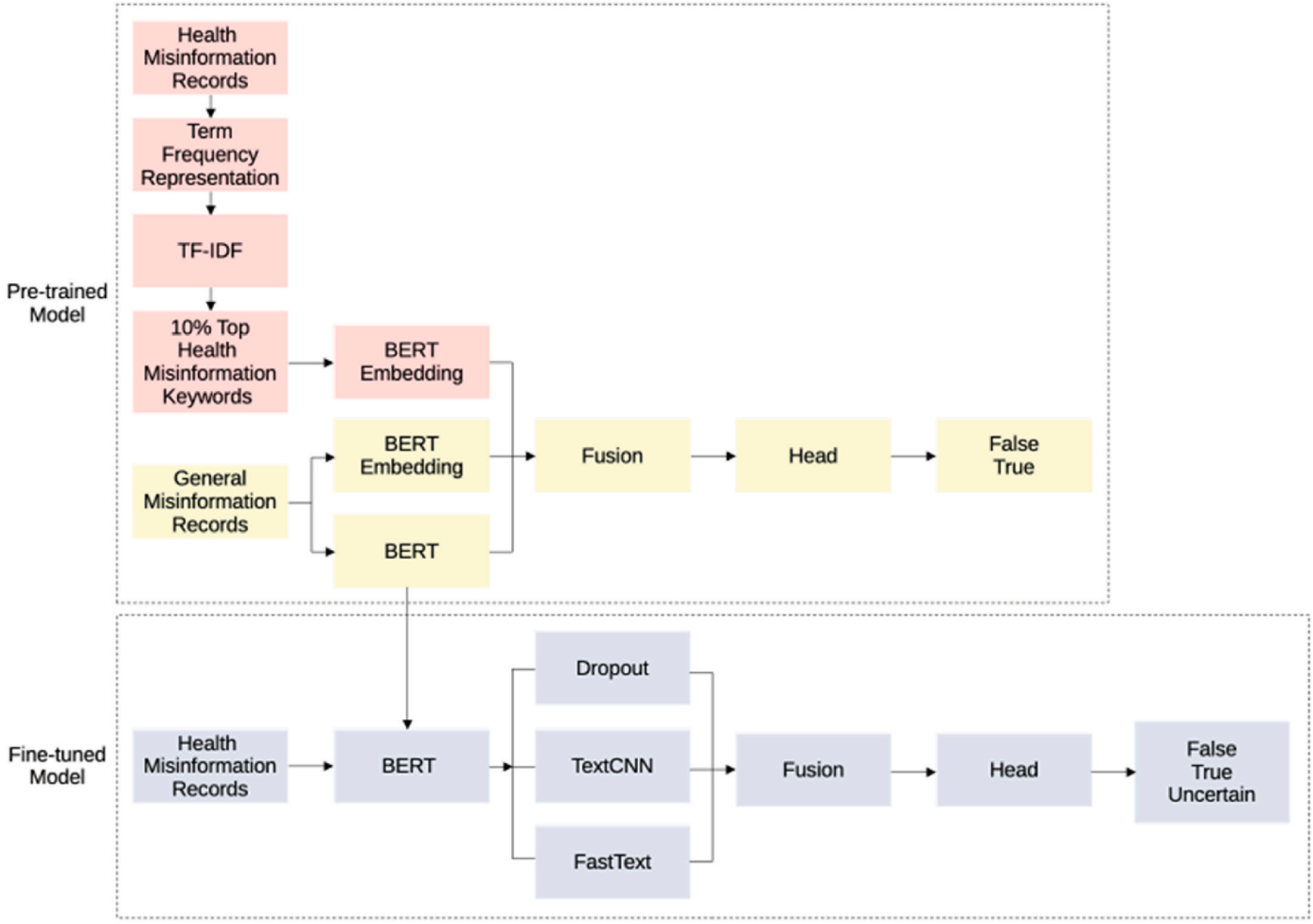

Fig. 1. Hybrid architecture of the proposed model.

1. 希腊红灯区重开后嫖客只能在房间呆15分钟 0
   After the re-open of red-light strict in Grace, customers could only stay 15 min.
2. 世界男高音多明戈因新冠肺炎于3月26日去世 0
   The world renowned tenor Placido Domingo passed away on March 26 due to COVID-19 related complications.
3. 新冠肺炎男性患者的死亡率更高 0
   The death rate of male COVID-19 patients is higher than female.
4. 武汉三口之家迪士尼游玩确诊新型冠状病毒肺炎 0
   A family from Wuhan who was staying in Shanghai Disneyland have been confirmed to be infected by COVID-19.
5. 因疫情原因，中国驻纳米比亚使馆组织包机回国 0
   Due to the spread of COVID-19, the Chinese Embassy in Namibia organised the local Chinese to return to China by Air charter.

6. 深呼吸后屏气5-10秒，如果没咳嗽，说明没有肺炎 1
   Hold your breath for 5-10 seconds after taking a deep breath. If you do not cough, it means you did not catch pneumonia.
7. 研究发现两成新冠患者会在3个月内患上精神病 1
   New research finds that 20% of COVID-19 patients will get mental diseases within 3 months.
8. 新型冠状病毒在2018年就被发现了 1
   The novel coronavirus was discovered in 2018.
9. 二氧化氯稀释后放进加湿器，可以有效预防新冠肺炎 1
   It is an efficient way to prevent COVID-19 by diluting chlorine dioxide and putting it into a humidifier.
10. 喝红酒能抵抗新型冠状病毒，延缓病情发展 1
    Drinking red wine can help resist COVID-19 and delay the development of the disease.

11. 家庭成员建议不共用毛巾 2
    Family members are not suggested to share towels.
12. 儿童在做体育运动或体力活动时，不应戴口罩，以免影响呼吸 2
    Children should not wear masks when they are doing sports or physical activities, so as not to block their breathing.
13. 判断人体是否感染了新冠病毒，常用的检测方法试核酸检测和血清学抗体检测 2
    The PCR and IgG/IgM tests are frequent used methods to check if someone is infected by the COVID-19.
14. 任何年龄的任何人都可能因新冠病毒而生病、病重或者死亡 2
    Anyone at any age can get sick, become seriously ill, or die from the COVID-19.
15. 乘坐公交车、地铁出行，必须全程正确佩戴口罩 2
    It is necessary always to wear the mask properly when taking bus or metro.

Fig. 2. Representative examples with their labels from the Chinese health misinformation dataset, with English translations shown below.

each dataset is randomly selected, and the remaining 90 % is randomly split into training and testing sets at a 3:1 ratio, with detailed information provided in Table 1. For further details, please refer to (Luo et al., 2024) (Luo et al., 2021) (Patwa et al., 2021).

Regarding general misinformation records, the Chinese dataset originates from the Biendata competition for fake news detection (https://www.heywhale.com/mw/dataset/5e096a422823a10036b10a81). After removing low-quality entries, it consists of 38,455 records, with 19,170 labeled as false and 19,284 labeled as true. In contrast, the English dataset is derived from the Kaggle fake news detection dataset https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets), where each record

```
1. We're building out some slack channels for different types of @COVID19Tracking users.        0
2. Colombia is the country with less cases and deaths for coronavirus per inhabitant in America.        0
3. A lab in Thane was closed for reporting false positive cases of COVID-19     0
4. Schedules for the sale of milk newspapers grocery and medical stores have been issued by the Mumbai Police commissioner.     0
5. Democratic primary candidates find common ground in being high risk for coronavirus  0

6. Coronavirus sticks faster on men with shaved heads.  1
7. Health officials advise everyone to follow random Facebook profile for all of your coronavirus information #CoronavirusOutbreak      1
8. The coronavirus was created in a lab and patented.   1
9. There has been no death due to COVID-19 in Israel as they mix lemon and baking soda in their tea. This combination kills coronavirus.        1
10.Bill Gates admits his COVID-19 vaccine might kill nearly 1 million people.   1

11.Many people could not identify having contact with a person who had #COVID19. Take steps to protect yourself and slow the spread.    2
12.flu and coronavirus collide we're in for real trouble.       2
13.Scale-up of testing is key in assessing the #COVID19Nigeria situation and coordinating response across the country.  2
14.Please convince your patients to wear masks out in public.   2
15.If testing and tracing is done quickly and completely it can contain COVID without the need for costly lockdowns.    2
```

Fig. 3. Representative examples with their labels from the English health misinformation dataset.

Table 1
Class distribution of Chinese and English health misinformation datasets.

		0	1	2	Total
Chinese dataset	Training set	288	191	234	713
	Validation set	41	26	38	105
	Test set	106	64	67	237
	Total	435	281	339	1055
English dataset	Training set	576	554	551	1681
	Validation set	75	92	82	249
	Test set	179	184	197	560
	Total	830	830	830	2490

contains both a title and text. Since the text portion in the true group is generally longer and more formal than that in the false group, only the title portion was selected for this study. Additionally, many titles in the false group have each word capitalized, while the true group does not follow this pattern. To ensure consistency, titles in both groups were preprocessed by converting all uppercase text to lowercase. Furthermore, some titles in the false group end with bracketed tags such as (VIDEO), (IMAGE), or (TWEET), which were removed during preprocessing. After filtering out low-quality records, the English dataset contains 42,986 records, with 21,569 labeled as false and 21,417 labeled as true. The two general misinformation datasets were partitioned into training, validation, and test sets following the same procedure as the health misinformation datasets. In addition, the Chinese and English health misinformation datasets, together with the processed Chinese and English general misinformation datasets, are available in (https://www.dropbox.com/scl/fo/lytsr6mitj6gevil0hz04/ACU52OThycRJsAy0y_PYVtw?rlkey=pds17om28lub7slxzmt0ooy0n&st=yobmt2hv&dl=0).

4.1.3. Evaluation metrics

Precision, recall, F1-score, and accuracy are the evaluation metrics used to assess the performance of both the hybrid transfer learning model and the comparison models.

Precision:Precision for class i is the ratio of correctly predicted instances of class i to the total instances predicted as class i. The precision for each class and the macro-averaged precision are formulated as:

$$P_i = \frac{T_{ii}}{T_{ii} + \sum_{j \neq i} T_{ji}} \tag{18}$$

$$P_{macro} = \frac{1}{n} \sum_i P_i \tag{19}$$

Recall: Recall for class i is the ratio of correctly predicted instances of class i to the total instances that actually belong to class i. The recall for each class and the macro-averaged recall are formulated as:

$$R_i = \frac{T_{ii}}{T_{ii} + \sum_{j \neq i} T_{ij}} \tag{20}$$

$$R_{macro} = \frac{1}{n} \sum_i R_i \tag{21}$$

F1-score: It is the harmonic mean of precision and recall, providing a balance between the two metrics. The F1-score for each class and the macro-averaged F1-score are formulated as:

$$F1_i = 2 \times \frac{P_i \times R_i}{P_i + R_i} \tag{22}$$

$$F1_{macro} = \frac{1}{n} \sum_i F1_i \tag{23}$$

Accuracy: It is the ratio of correctly predicted instances to the total number of instances and is formulated as:

$$ACC = \frac{\sum_i T_{ii}}{\sum_{ij} T_{ij}} \tag{24}$$

4.2. Results

4.2.1. Comparison with eight commonly employed deep learning models

The performance of the proposed hybrid transfer learning model for ternary classification in health misinformation detection is compared with eight widely used deep learning models (Luo et al., 2024) using P_{macro}, R_{macro}, $F1_{macro}$ and ACC. These models include fastText, three Recurrent Neural Network (RNN)-based models, two Convolutional Neural Network (CNN)-based models, and two transformer-based models. Specifically, the RNN-based models are TextRNN (Liu et al., 2016), TextRNN_Att (Zhou et al., 2016), and TextRCNN (Lai et al., 2015); the CNN-based models are TextCNN and DPCNN (Johnson and Zhang, 2017); and the transformer-based models are Transformer (Vaswani et al., 2017) and BERT. The test results for Chinese records are presented in Table 2, while those for English records are shown in Table 3.

In the Chinese dataset, the hybrid transfer learning model achieves the highest performance across all evaluation metrics. Specifically, it records a P_{macro} of 0.8710, R_{macro} of 0.8724, $F1_{macro}$ of 0.8693, and ACC of 0.8776, surpassing the other models, including BERT, which is the

Table 2
Test results for ternary classification in health misinformation detection in Chinese.

	P_{macro}	R_{macro}	$F1_{macro}$	ACC
FastText	0.7652	0.7673	0.7655	0.7806
TextRNN	0.7119	0.7174	0.7109	0.7173
TextRNN_Att	0.7417	0.7491	0.7441	0.7553
TextRCNN	0.7229	0.7208	0.7204	0.7342
TextCNN	0.7683	0.7696	0.7689	0.7806
DPCNN	0.6993	0.7002	0.6972	0.7173
Transformer	0.6595	0.6708	0.6618	0.6709
BERT	0.8120	0.8136	0.8075	0.8101
Hybrid	**0.8710**	**0.8724**	**0.8693**	**0.8776**

Table 3
Test results for ternary classification in health misinformation detection in English.

	P_{macro}	R_{macro}	$F1_{macro}$	ACC
FastText	0.6571	0.6551	0.6482	0.6589
TextRNN	0.6236	0.6239	0.6205	0.6250
TextRNN_Att	0.6519	0.6471	0.6409	0.6500
TextRCNN	0.6351	0.6376	0.6300	0.6411
TextCNN	0.6562	0.6556	0.6557	0.6571
DPCNN	0.6138	0.6114	0.6107	0.6143
Transformer	0.5461	0.5469	0.5452	0.5500
BERT	0.7594	0.7527	0.7481	0.7571
Hybrid	**0.7904**	**0.7889**	**0.7885**	**0.7911**

second-best performer with an $F1_{macro}$ of 0.8075 and ACC of 0.8101. FastText and TextCNN also perform moderately well, with $F1_{macro}$ values of 0.7655 and 0.7689, respectively. However, other models lag significantly behind, highlighting their limitations in this context. In the English dataset, the hybrid transfer learning model once again outperforms the other models, recording a P_{macro} of 0.7904, R_{macro} of 0.7889, $F1_{macro}$ of 0.7885, and ACC of 0.7911, demonstrating its robustness across languages. BERT remains the second-best performer with an $F1_{macro}$ of 0.7481 and ACC of 0.7571, but the performance gap with the hybrid model is slightly narrower compared to the Chinese dataset. FastText and TextCNN perform similarly, with $F1_{macro}$ values of 0.6482 and 0.6557, respectively. However, as in the Chinese dataset, other models yield weaker results.

The hybrid model consistently outperforms the other deep learning models on both Chinese and English datasets, achieving accuracy improvements of at least 6.75 % and 3.4 %, respectively. This demonstrates its strong capabilities in detecting health misinformation across different languages. Moreover, since the amount of health misinformation data is relatively limited, models with pre-trained embeddings or simpler architectures tend to outperform more complex models. The hybrid model, which effectively combines transfer learning and simpler model structures, leverages the benefits of pre-trained knowledge while avoiding overfitting.

4.2.2. Ablation study

To quantify the contribution of each component and understand how multi-modal fusion and model enhancements affect detection performance, an ablation study of the proposed hybrid model is conducted. The effects of removing or retaining the attention mechanism and pretraining, as well as combinations of different feature extractors, are systematically evaluated. The test results for Chinese records are presented in Table 4, while those for English records are shown in Table 5.

Table 4
Ablation study results for the hybrid model in health misinformation detection in Chinese.

	P_{macro}	R_{macro}	$F1_{macro}$	ACC
Hybrid	0.8710	0.8724	0.8693	0.8776
BERT + TextCNN + fastText (w/o Attention, with Pretraining)	0.8369	0.8423	0.8368	0.8481
BERT + TextCNN + fastText (w/o Pretraining, with Attention)	0.7904	0.7864	0.7832	0.7890
BERT + TextCNN (with Attention & Pretraining)	0.8504	0.8615	0.8525	0.8565
BERT + fastText (with Attention & Pretraining)	0.8456	0.8566	0.8491	0.8565
TextCNN + fastText (with Attention & Pretraining)	0.8247	0.8403	0.8266	0.8312
BERT only (with Attention & Pretraining)	0.8308	0.8367	0.8332	0.8397
TextCNN only (with Attention & Pretraining)	0.8139	0.8259	0.8123	0.8143
fastText only (with Attention & Pretraining)	0.8433	0.8459	0.8438	0.8523

Table 5
Ablation study results for the hybrid model in health misinformation detection in English.

	P_{macro}	R_{macro}	$F1_{macro}$	ACC
Hybrid	0.7904	0.7889	0.7885	0.7911
BERT + TextCNN + fastText (w/o Attention, with Pretraining)	0.7690	0.7680	0.7665	0.7714
BERT + TextCNN + fastText (w/o Pretraining, with Attention)	0.7703	0.7438	0.7366	0.7482
BERT + TextCNN (with Attention & Pretraining)	0.7678	0.7662	0.7655	0.7696
BERT + fastText (with Attention & Pretraining)	0.7815	0.7772	0.7754	0.7804
TextCNN + fastText (with Attention & Pretraining)	0.7798	0.7794	0.7786	0.7821
BERT only (with Attention & Pretraining)	0.7714	0.7712	0.7698	0.7750
TextCNN only (with Attention & Pretraining)	0.7769	0.7748	0.7743	0.7768
fastText only (with Attention & Pretraining)	0.7864	0.7845	0.7835	0.7875

For the Chinese dataset, the highest performance is obtained by the hybrid model. A performance drop is observed when the attention mechanism is removed while pretraining is retained ($F1_{macro} = 0.8368$), and a larger decrease occurs when pretraining is excluded while attention is maintained ($F1_{macro} = 0.7832$), indicating that both components are critical for Chinese text processing. Among single-modality models, fastText achieves the highest $F1_{macro}$ (0.8438), followed by BERT (0.8332) and TextCNN (0.8123). Dual-modality combinations consistently outperform single models, demonstrating the effectiveness of multi-modal feature fusion. For the English dataset, the highest metrics are also achieved by the hybrid model, although the improvement over the strongest single-modality model, fastText ($F1_{macro} = 0.7835$), is modest. Performance decreases are observed when attention or pretraining is ablated, but the magnitude is smaller than in Chinese. Dual-modality combinations also produce improvements over single models, indicating that multi-modal fusion remains beneficial.

Across both the Chinese and English datasets, the ablation study consistently shows that the hybrid model achieves the best performance when both attention and pretraining are included, which makes it clear that multi-modal fusion is broadly effective. Single-modality and dual-modality models fall behind, which confirms that combining several feature extractors is important for strong detection across languages. Moreover, performance drops most substantially when pretraining is excluded, underscoring its critical importance. This effect is particularly evident in the Chinese dataset, likely due to the higher quality of the Chinese pretraining data that enables the model to build richer representations. Incorporating a higher-quality English general misinformation dataset may further enhance performance on English records.

4.2.3. Comparison with fastText, TextCNN, and BERT for each class

Since the hybrid model is developed by combining BERT, TextCNN, and fastText, and these models are ranked among the top four in subsection 4.2.1, the performance of the proposed hybrid transfer learning model for ternary classification in health misinformation detection is further compared with BERT, TextCNN, and fastText using P_i, R_i, and $F1_i$ for the uncertain, false, and true classes, respectively. The test results for Chinese records are presented in Table 6, while those for English records are shown in Table 7.

Concerning the Chinese records, the hybrid model shows strong gains across all three classes. For class 0, P_i and R_i reach 0.9223 and 0.8962, resulting in an $F1_i$ score of 0.9091. For class 1, the $F1_i$ score rises to 0.8099, compared with 0.6774 for fastText, 0.6875 for TextCNN, and 0.7123 for BERT, representing the largest improvement among all classes. For class 2, R_i reaches 0.9552 and the $F1_i$ score is 0.8889, confirming the hybrid model's robust ability to capture this category. BERT performs reasonably well on classes 0 and 2, with $F1_i$ scores of 0.8485

Table 6
Test results for each class in ternary classification of health misinformation detection in Chinese.

		0	1	2
FastText	P_i	0.8558	0.7000	0.7397
	R_i	0.8396	0.6562	0.8060
	$F1_i$	0.8476	0.6774	0.7714
	ACC	0.7806		
TextCNN	P_i	0.8381	0.6875	0.7794
	R_i	0.8302	0.6875	0.7910
	$F1_i$	0.8341	0.6875	0.7852
	ACC	0.7806		
BERT	P_i	0.9130	0.6341	**0.8889**
	R_i	0.7925	**0.8215**	0.8358
	$F1_i$	0.8485	0.7123	0.8615
	ACC	0.8101		
Hybrid	P_i	**0.9223**	**0.8596**	0.8312
	R_i	**0.8962**	0.7656	**0.9552**
	$F1_i$	**0.9091**	**0.8099**	**0.8889**
	ACC	**0.8776**		

Table 7
Test results for each class in ternary classification of health misinformation detection in English.

		0	1	2
FastText	P_i	0.6429	0.6396	0.6887
	R_i	0.4525	0.7717	0.7411
	$F1_i$	0.5311	0.6995	0.7139
	ACC	0.6589		
TextCNN	P_i	0.5989	0.6554	0.7143
	R_i	0.6257	0.6304	0.7107
	$F1_i$	0.6120	0.6427	0.7125
	ACC	0.6571		
BERT	P_i	**0.7752**	0.7150	0.7880
	R_i	0.5587	**0.8315**	**0.8680**
	$F1_i$	0.6494	0.7688	0.8261
	ACC	0.7571		
Hybrid	P_i	0.7622	**0.7438**	**0.8653**
	R_i	**0.6983**	0.8207	0.8477
	$F1_i$	**0.7289**	**0.7804**	**0.8564**
	ACC	**0.7911**		

and 0.8615, but falls behind the hybrid model, particularly for class 1. FastText shows relatively weaker performance overall, while TextCNN performs slightly better than fastText but still lags behind BERT. Regarding the English records, the hybrid model again delivers the strongest performance across the three classes. In class 0, the $F1_i$ score reaches 0.7289, surpassing fastText (0.5311), TextCNN (0.6120), and BERT (0.6494). Class 1 shows an $F1_i$ of 0.7804, which exceeds fastText (0.6995), TextCNN (0.6427), and BERT (0.7688). Class 2 stands out, with the hybrid model achieving P_i of 0.8653, R_i of 0.8477, and an $F1_i$ score of 0.8564, clearly ahead of fastText ($F1_i = 0.7139$), TextCNN ($F1_i = 0.7125$), and BERT ($F1_i = 0.8261$). While BERT maintains high R_i for classes 1 and 2, its lower P_i results in less balanced performance. Among the simpler models, fastText tends to perform slightly better on class 2, whereas TextCNN provides more consistent but moderate results across all classes.

In both datasets, the hybrid model's ability to combine the strengths of BERT, TextCNN, and fastText allows it to deliver better performance in detecting health misinformation across multiple classes. This combination outperforms any individual model, providing a more balanced and robust outcome. Moreover, the hybrid model shows the largest improvement in class 1 for the Chinese dataset ($F1_i$ increase of 0.0976) and in class 0 for the English dataset ($F1_i$ increase of 0.0795). These classes are originally the weakest in the other three models, indicating that the hybrid approach is particularly effective at enhancing underperforming categories.

4.2.4. Confusion matrix

The confusion matrix is a valuable tool for visualizing the model's classification performance by providing a detailed comparison between the predicted and actual classes. In health misinformation detection, it underscores the model's ability to accurately classify uncertain, false, and true categories, while also highlighting any misclassifications. Fig. 4 presents the confusion matrix for the hybrid model, compared with BERT, TextCNN, and fastText, based on Chinese records, while Fig. 5 displays the results for English records.

In the Chinese dataset, the hybrid model outperforms all others, performing particularly well in class 2 (96 %) and class 0 (90 %). BERT also achieves good results, though it shows notable misclassifications in class 0, where 19 instances are incorrectly labeled as false. TextCNN and fastText demonstrate relatively weaker performance, especially in class 1, with misclassification rates of 31 % and 34 %, respectively. In the English dataset, the hybrid model again performs best, maintaining strong results for class 1 (82 %) and class 2 (85 %), but with more misclassifications compared to the Chinese dataset. BERT exhibits a high number of misclassifications in class 0, with many instances incorrectly predicted as false. TextCNN and fastText perform moderately, while fastText shows the weakest performance, particularly in class 0, where 55 % of instances are misclassified.

The confusion matrix further confirms that the hybrid model generally outperforms the other individual models across both datasets, demonstrating higher correct classifications and fewer misclassifications. In addition, the Chinese results are more polarized, with class 2 nearly perfectly classified while class 1 lags at 77 %, whereas the English results are more balanced across classes (70–85 %). This pattern may be partly explained by data distribution since the Chinese dataset is generally balanced but class 1 has fewer examples than the other classes, which may limit the model's ability to capture its distinguishing features. In contrast, the English dataset is evenly distributed across the three classes, providing fully balanced training signals and yielding more consistent performance. Therefore, strategies such as data augmentation, class-weighted loss functions, or targeted oversampling for class 1 in the Chinese dataset could be considered to improve overall performance.

5. Conclusions and future works

This paper introduces a hybrid transfer learning model tailored for ternary classification in health misinformation detection. The model begins by leveraging a pre-trained network to classify general misinformation as either false or true, while integrating health-related keywords into the dataset. It then refines the classification process by combining BERT, TextCNN, and fastText with an attention mechanism to categorize health misinformation into uncertain, false, or true classes. Firstly, the model's performance was evaluated against eight widely used deep learning models. It consistently achieved the best results in both Chinese and English datasets, demonstrating strong capabilities in detecting health misinformation across different languages. Secondly, an ablation study was conducted to quantify the contribution of each component, showing that multi-modal fusion was broadly effective and highlighting the critical importance of pretraining. Thirdly, the hybrid model was compared with BERT, TextCNN, and fastText for the uncertain, false, and true classes. The results emphasized its ability to improve classification accuracy across multiple categories, particularly the effectiveness in enhancing underperforming classes. Finally, a confusion matrix analysis highlighted the relationship between variations in model performance and the underlying data distribution in the two datasets.

In the future, enhancing the quality of the English general misinformation dataset and increasing class samples in the Chinese health misinformation dataset to achieve a more balanced distribution should be considered. At present, both the Chinese and English datasets have limitations, and improving them could further enhance model performance. Moreover, additional post-hoc analysis methods may be required

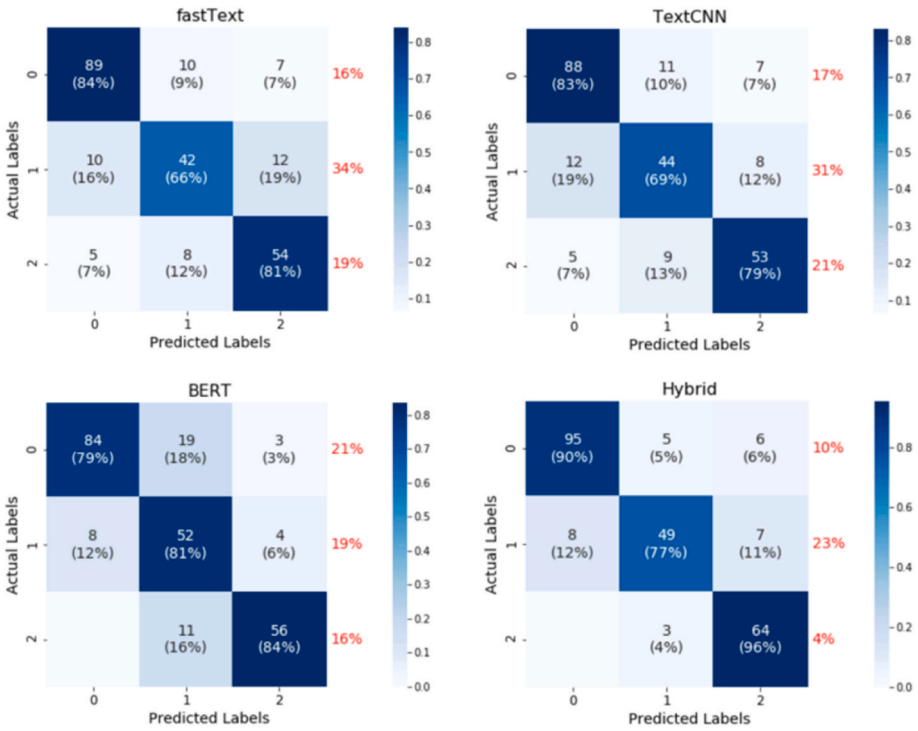

Fig. 4. Confusion matrix for ternary classification in health misinformation detection in Chinese.

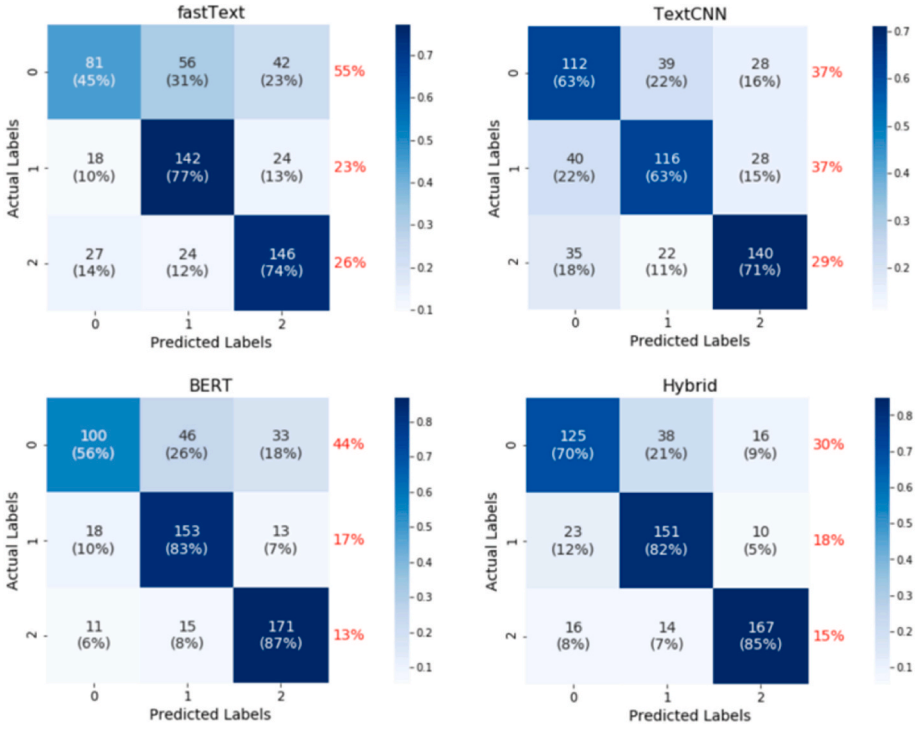

Fig. 5. Confusion matrix for ternary classification in health misinformation detection in English.

to better understand the decision-making process. While the hybrid model offers partial interpretability through component contribution assessment, it remains largely a black-box model due to its complexity. Finally, incorporating multimodal data and developing multimodal approaches will be explored. As the current method deals only with text, leveraging additional modalities could further improve classification performance and provide richer information for model interpretation.

CRediT authorship contribution statement

Jia Luo: Writing – original draft, Methodology, Conceptualization. **Yang Yang:** Writing – review & editing, Validation. **Xiaoye Feng:** Validation, Data curation. **Didier El Baz:** Writing – review & editing, Validation, Supervision.

Informed consent

This work does not contain any studies with human participants performed by any of the authors.

Ethical approval

Not applicable.

Funding

This work is supported by the Beijing Natural Science Foundation (Grant No. 9242003), the Natural Science Foundation of Chongqing, China (Grant No. CSTB2023NSCQ-MSX0391), and the National Natural Science Foundation of China (Grant No. 72104016).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this study are available in (https://www.dropbox.com/scl/fo/lytsr6mitj6gevil0hz04/ACU52OThycRJsAy0y_PYVtw?rlkey=pds17om28lub7slxzmt0ooy0n&st=yobmt2hv&dl=0). Additional data and supporting materials are available from the first and corresponding author upon request.

References

Alghamdi, J., Lin, Y., Luo, S., 2023. Towards COVID-19 fake news detection using transformer-based models. Knowl. Base Syst. 274, 110642.
Alghamdi, J., Luo, S., Lin, Y., 2024a. A comprehensive survey on machine learning approaches for fake news detection. Multimed. Tool. Appl. 83 (17), 51009–51067.
Alghamdi, J., Lin, Y., Luo, S., 2024b. Unveiling the hidden patterns: a novel semantic deep learning approach to fake news detection on social media. Eng. Appl. Artif. Intell. 137, 109240.
Alghamdi, J., Lin, Y., Luo, S., 2024c. Enhancing hierarchical attention networks with CNN and stylistic features for fake news detection. Expert Syst. Appl. 257, 125024.
Arunthavachelvan, K., Raza, S., Ding, C., 2024. A deep neural network approach for fake news detection using linguistic and psychological features. User Model. User-Adapted Interact. 34 (4), 1043–1070.
Bird, S., Klein, E., 2006. Regular Expressions for Natural Language Processing. University of Pennsylvania.
Bondielli, A., Marcelloni, F., 2019. A survey on fake news and rumour detection techniques. Inf. Sci. 497, 38–55.
Chen, M.Y., Lai, Y.W., Lian, J.W., 2023. Using deep learning models to detect fake news about COVID-19. ACM Trans. Internet Technol. 23 (2), 1–23.
Chen, X., Jian, Y., Ke, L., Qiu, Y., Chen, X., Song, Y., Wang, H., 2024a. A deep semantic-aware approach for Cantonese rumor detection in social networks with graph convolutional network. Expert Syst. Appl. 245, 123007.
Chen, Z., Zhuang, F., Liao, L., Jia, M., Li, J., Huang, H., 2024b. A syntactic multi-level interaction network for rumor detection. Neural Comput. Appl. 36 (4), 1713–1726.
Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., et al., 2021. A COVID-19 rumor dataset. Front. Psychol. 12, 644801.
Chou, W.S., Gaysynsky, A., Vanderpool, R.C., 2018. The role of health communication in the spread of health misinformation. Am. J. Publ. Health 108 (5), 601–602.
Cinelli, M., Quattrociocchi, W., Galeazzi, A., et al., 2020. The COVID-19 social media infodemic. Sci. Rep. 10, 16598.
Dharawat, A., Lourentzou, I., Morales, A., Zhai, C., 2022. Drink bleach or do what now? Covid-hera: a study of risk-informed health decision making in the presence of covid-19 misinformation. Proceed. Int. AAAI Conference Web Social Media 16, 1218–1227.
Di Sotto, S., Viviani, M., 2022. Health misinformation detection in the social web: an overview and a data science approach. Int. J. Environ. Res. Publ. Health 19 (4), 2173.
FactCheck.org. About us. https://www.factcheck.org.
Friggeri, A., Gallus, S., Adamic, L.A., 2014. Rumor Cascades. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM), pp. 101–110.
Hajek, P., Sahut, J.M., Munk, M., Munkova, D., 2024. Detecting fake news on social networks via linguistic features and information-seeking patterns during the Covid-19 period. Ann. Oper. Res. 1–24.
Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T., 2021. ArCOV19-rumors: arabic COVID-19 Twitter dataset for misinformation detection. In: Proceedings of the 6th Arabic Natural Language Processing Workshop. pp. 72-81.
Health Feedback. How we review claims. https://healthfeedback.org.
Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (Volume 1:Long and Short Papers), pp. 4171–4186.
https://www.heywhale.com/mw/dataset/5e096a422823a10036b10a81.
Hussain, A., Ali, W., Ahmad, A., Iqbal, M.S., Moqurrab, S.A., Paul, A., et al., 2025. Enhancing COVID-19 misinformation detection through novel attention mechanisms in NLP. Expert Syst. 42 (1), e13571.
Johnson, R., Zhang, T., 2017. Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 562–570.
Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2017. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 427–431.
Kim, Y., 2014. Convolutional neural networks for sentence classification. In: In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1746–1751.
Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., Choi, D., 2021. FibVID: comprehensive fake news diffusion dataset during the COVID-19 period. Telematics Inf. 64, 101688.
Lai, S., Xu, L., Liu, K., Zhao, J., 2015. Recurrent convolutional neural networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29, No. 1.
Liu, P., Qiu, X., Huang, X., 2016. Recurrent neural network for text classification with multi-task learning. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 2873–2879.
Liu, Y., Yu, K., Wu, X., Qing, L., Peng, Y., 2019. Analysis and detection of health-related misinformation on Chinese social media. IEEE Access 7, 154480–154489.
Luo, J., Xue, R., Hu, J., El Baz, D., 2021. Combating the infodemic: a chinese infodemic dataset for misinformation identification. In: Healthcare. MDPI, p. 1094. Vol. 9, No. 9.
Luo, J., Peng, D., Shi, L., El Baz, D., Liu, X., 2023. A comparative analysis of the COVID-19 Infodemic in English and Chinese: insights from social media textual data. Front. Public Health 11, 1281259.
Luo, J., El Baz, D., Shi, L., 2024. Utilizing deep learning models for ternary classification in COVID-19 infodemic detection. Digital Health 10, 1–12.
Murayama, T., 2021. Dataset of fake news detection and fact verification: a survey. arXiv preprint arXiv:2111.03299.
Nguyen, T., Shirai, K., 2015. Hybrid approaches to text classification. Inf. Process. Manag. 51 (4), 563–571.
Ozcelik, O., Toraman, C., Can, F., 2025. Detecting misinformation on social media using community insights and contrastive learning. ACM Transac. Intelligent System. Technol. 16 (2), 1–27.
Papanikou, V., Papadakos, P., Karamanidou, T., Stavropoulos, T.G., Pitoura, E., Tsaparas, P., 2024. Health misinformation in social networks: a survey of IT approaches. arXiv preprint arXiv:2410.18670.
Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M.S., et al., 2021. Fighting an infodemic: Covid-19 fake news dataset. In combating online hostile posts in regional languages during emergency situation. In: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1. Springer International Publishing, pp. 21–29.
Pennycook, G., Rand, D.G., 2018. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc. Natl. Acad. Sci. 115 (3), 576–583.
Praseed, A., Rodrigues, J., Thilagam, P.S., 2023. Hindi fake news detection using transformer ensembles. Eng. Appl. Artif. Intell. 119, 105731.
Raja, E., Soni, B., Borgohain, S.K., 2024. Fake news detection in Dravidian languages using multiscale residual CNN_BiLSTM hybrid model. Expert Syst. Appl. 250, 123967.
Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for IDF. J. Doc. 60 (5), 503–520.
Schlicht, I.B., Fernandez, E., Chulvi, B., Rosso, P., 2024. Automatic detection of health misinformation: a systematic review. J. Ambient Intell. Hum. Comput. 15 (3), 2009–2021.
Shahi, G.K., Dirkson, A., Majchrzak, T.A., 2021. An Exploratory Study of COVID-19 Misinformation on Twitter, vol. 22. Online Social Networks and Media, 100104.
Shang, L., Zhang, Y., Chen, B., Zong, R., Yue, Z., Zeng, H., et al., 2024. MMAdapt: a knowledge-guided multi-source multi-class domain adaptive framework for early health misinformation detection. In: Proceedings of the ACM on Web Conference 2024, pp. 4653–4663.
Sicilia, R., Giudice, S.L., Pei, Y., Pechenizkiy, M., Soda, P., 2018. Twitter rumour detection in the health domain. Expert Syst. Appl. 110, 33–40.
Sicilia, R., Merone, M., Valenti, R., Soda, P., 2021. Rule-based space characterization for rumour detection in health. Eng. Appl. Artif. Intell. 105, 104389.
Swire-Thompson, B., Lazer, D., 2020. Public health and online misinformation: challenges and recommendations. Annu. Rev. Publ. Health 41, 433–451.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science 359 (6380), 1146–1151.

Xia, H., Wang, Y., Zhang, J.Z., Zheng, L.J., Kamal, M.M., Arya, V., 2023. COVID-19 fake news detection: a hybrid CNN-BiLSTM-AM model. Technol. Forecast. Soc. Change 195, 122746.
Zhao, Y., Da, J., Yan, J., 2021. Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches. Inf. Process. Manag. 58 (1), 102390.
Zhou, X., Zafarani, R., 2020. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv. 53 (5), 1–40.
Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B., 2016. Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 207–212.