Displays 90 (2025) 103154

Contents lists available at ScienceDirect

Displays

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

A joint learning framework for fake news detection ☆

Muhammad Abdullah [a], Zan Hongying [a,*], Arifa Javed [a], Orken Mamyrbayev [b], Fabio Caraffini [c,*], Hassan Eshkiki [c]

[a] School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, Henan, China
[b] Institute of Information and Computational Technologies, Almaty, Kazakhstan
[c] Department of Computer Science, Swansea University, Swansea SA1 8EN, UK

ARTICLE INFO

Keywords:
Joint learning
BERT
Semantics
NLP
Fake news
RFC
NER

ABSTRACT

This paper presents a joint learning framework for fake news detection, introducing an Enhanced BERT model that integrates named entity recognition, relational feature classification, and Stance Detection through a unified multi-task approach. The model incorporates task-specific masking and hierarchical attention mechanisms to capture both fine-grained and high-level contextual relationships across headlines and body text. Cross-task consistency losses are applied to ensure coherence and alignment with external factual knowledge. We analyse the average distance from components to the centroid of a news sample to differentiate genuine information from falsehoods in large-scale text data effectively. Experiments on two FakeNewsNet datasets show that our framework outperforms state-of-the-art models, with accuracy improvements of 2.17% and 1.03%. These results indicate the potential for applications needing detailed text processing, like automatic summarisation and misinformation detection.

1. Introduction

The widespread dissemination of misinformation represents a profound challenge in the contemporary digital era, where fake news easily and quickly reaches a larger audience through social media. Misleading narratives negatively impact social structures by interfering with main events, for example, political and humanitarian [1,2], and affect democratic ideals [3].

To fight this trend, numerous fact-checking organisations [4], such as *Snopes.com*, *FactCheck.org*, *PolitiFact.com*, allocate substantial resources to mitigate misinformation. However, their approach, which relies mainly on domain experts to manually verify the accuracy of news stories, cannot cope with the current volume and rate of fake news. In this light, developing accurate automated fake news detection systems has become a necessity. Researchers are currently making an effort to achieve this goal; see, e.g. [2,5,6].

The first strategies adopted to implement automatic fake news detection involve manually preparing large sets of features based on various aspects, such as the content of the news-incorporating linguistic features like syntax, grammar, and word usage-user profiles, and the paths of news propagation [7,8]. These features are then used to train Machine Learning (ML) classifiers to assess the authenticity of the news [9,10]. Certain studies have facilitated the identification of significant network-based and user-profile features, as evidenced in [11], which employed the XG Boost model to detect deceptive behaviour on social networks. However, with these ML approaches, it is hard to cover all the linguistic nuances in fake news, which usually spans diverse topics, styles, and platforms [12]. They are also less effective when dealing with well-written or complex fake news and require contextual data to improve effectiveness. In addition, classic ML models struggle with the evolving tactics of misinformation and the increasing complexity of digital communication [13], making them unsuitable for detecting intricate linguistic and contextual patterns [14].

Deep learning (DL) addresses these limitations by autonomously learning to differentiate patterns within news content and propagation paths [13,15]. However, standard deep learning models face several challenges in fake news detection. One key challenge is handling long-form news articles, where traditional models struggle due to issues such as exploding gradients and inefficient feature extraction, leading to reduced accuracy [16–19]. Another challenge is capturing the contextual dependencies between entities and their relationships, as fake news often manipulates factual connections between individuals, organisations, and events [20]. Techniques such as gradient clipping and

☆ This paper was recommended for publication by Guangtao Zhai.

* Corresponding author.
E-mail addresses: abdullah@gs.zzu.edu.cn (M. Abdullah), iehyzan@zzu.edu.cn (Z. Hongying), arifa.javed@gs.zzu.edu.cn (A. Javed), morkenj@mail.ru (O. Mamyrbayev), fabio.caraffini@swansea.ac.uk (F. Caraffini), h.g.eshkiki@swansea.ac.uk (H. Eshkiki).

https://doi.org/10.1016/j.displa.2025.103154
Received 17 November 2024; Received in revised form 30 April 2025; Accepted 5 July 2025
Available online 23 July 2025

normalisation layers are commonly applied to mitigate some of these issues, yet they remain insufficient for ensuring high accuracy across diverse and complex misinformation patterns. While Recurrent Neural Networks (RNNs) effectively process sequential text data, they suffer from limitations in capturing long-term dependencies due to vanishing gradients. Similarly, Convolutional Neural Networks (CNNs), which are commonly used for text classification, struggle to model relationships across distant words because their reliance on local connectivity prevents them from fully understanding broader semantic structures [21]. The Extreme Fake News Detection (EFND) model represents a significant advancement by integrating textual, visual, and social contexts using multimodal factorised bilinear pooling and a multilayer perceptron for classification [12]. Despite these advancements, existing neural networks still struggle with misinformation detection because they fail to jointly consider the complex relationships among entities, textual inconsistencies, and stance variations within fake news articles [22]. The evolving nature of misinformation requires models that can analyse news holistically while efficiently handling large volumes of data [23].

This paper introduces a novel Joint Learning for Fake News Detection (JLFND) framework built by extending the popular pre-trained BERT [24], whose architecture is based on transformers and is well-suited for modelling fake news data. Precisely, our JLFND model enhances the BERT_large architecture by incorporating task-specific attention masking and a hierarchical attention mechanism, collectively improving the model's ability to analyse both local and global contextual features of news content. The incorporation of Named Entity Recognition (NER), Relational Feature Classification (RFC), and Stance Detection (SD) within an integrated multi-task framework renders the JLFND approach both comprehensive and robust for the detection of fake news. This methodology not only demonstrates high accuracy but also enhances the model's adaptability to the dynamic challenges presented by the evolving intricacies of misinformation.

The remainder of this article is organised as follows:

- Section 2 report advances and background information pertinent to this field;
- Section 3 describes the methodology of this study;
- Section 4 comments on the design process for the proposed fake news detection system;
- Section 5 describes the experimental setup;
- Section 6 presents and discusses the results;
- Section 7 concludes this work and highlights its contribution.

2. Related work

If DL represents a turning point for fake news detection, transformers are the methods that have completely revolutionised this field. Especially, BERT [24] is a milestone for most Natural Language Processing (NLP) contexts, including fake news detection [25]. The studies in [26,27] compared BERT with several ML models, including LSTM, BiLSTM, and CNN-BiLSTM, and confirmed its superiority in three large datasets, namely, CoAID, GossipCop, and PolitiFact. The results show that BERT outperforms the other models, achieving the highest accuracy and F1 scores across all datasets. Successors like Roberta and ALBERT feature optimised training methods to reduce model size for better performance [2]. Recent studies introduced models like FNDNS, which use transformer architecture to improve fake news detection and behaviour prediction [22]. Feature Gradient Method with Feature Regularisation Adversarial Training enhances robustness against new data [6]. Study [28] enhances misinformation detection on Twitter with content features, while [12,29] combined linguistic and social context for fake news classification. These methods depend on predefined features, reducing adaptability to new topics, styles, or platforms [30].

Contemporary methods make use of even more advanced models such as the HyproBERT model [15], which combines DistilBERT [31] for embedding and tokenisation with a CNN, BiGRU [32,33], and CapsNet layers [34]. This integration led to a significant improvement in accuracy and robustness. Researchers have explored auxiliary tasks such as NER to enhance fake news detection, as demonstrated by the study in [35], which shows that incorporating NER improves the model's ability to understand and classify content. Similarly, the study in [36] suggests that using relational characteristics significantly increases accuracy. These methods are effective but often require separate models or layers, increasing computational costs. Our proposed model, as outlined in Section 3, uses a streamlined approach with task-specific attention masking within BERT's architecture, eliminating the need for additional layers and separate task-specific encoders as seen in multi-component architectures like HyproBERT. This design lowers the model complexity, reducing computational demands while maintaining task flexibility and robustness in fake news detection.

Recent advancements in multi-modal fake news detection have focused on integrating diverse data sources such as text, images, and videos to improve classification accuracy. Wang et al. (2023) propose a transformer-based model for detecting fake news from positive unlabelled data by leveraging both textual and visual features [37]. Qian et al. (2021) introduce a hierarchical multi-modal approach that uses contextual attention mechanisms to enhance the detection of fake news by combining various data modalities [38]. Wu et al. (2021) present a co-attention network that simultaneously processes text and images, capturing the interactions between them for more effective detection [39]. Khattar et al. (2019) propose a variational autoencoder (VAE) model to integrate multiple modalities in a probabilistic framework, improving performance under uncertainty [40]. Finally, Wang et al. (2018) develop an event-adversarial neural network that integrates event-specific information with multi-modal data to better identify fake news [41]. Together, these works highlight the growing importance of multi-modal approaches in addressing the challenges of fake news detection.

Several Large Langue Models (LLMs) are gaining traction for the fake news detection task. In [42], the authors combine active learning techniques with pre-trained LLMs to reduce data requirements and computational costs while maintaining high detection accuracy. In [43], an LLM agent is designed to detect fake news. The Knowledge-Enhanced AutoPrompt (KEAP) method in [44] converts fake news detection into a prompt learning task with T5-generated templates. Using external entity knowledge, KEAP improves its ability to detect fake news through prompt learning and enriched context.

Noteworthy studies that deserve mention incorporate Transformers with Graph Convolutional Networks [45], employ the Graph Global Attention Network (GGAN) model [14] and other models like Similarity-Aware Fake News Detection (SAFE) [46], or even develop expainable systems such as use the Multifaceted Reasoning Network for Explainable Fake News Detection (MRE-EFND) [47] and the framework in [48], which features a graph attention mechanisms.

This examination of the recent literature shows that challenges that persist in fake news detection systems are attributable not only to the intricacies of patterns in long text data, but also to known issues in ML such as data imbalance. Current models often fail to capture the linguistic and contextual nuances needed for a reliable classification, as shown in the summary in Table 1.

3. Methodology

In ML jargon, the detection of fake news is a classic binary classification problem, i.e., with only 2 classes c_i ($i \in \{0, 1\}$), where each piece of news can be fake (C_0) or real (C_1). Given a dataset of news articles, the embeddings of the words in the articles are represented as vectors in a high-dimensional space, denoted as $\mathbf{d}$. The embeddings for a set of news articles can be organised into a matrix $\mathbf{N} \in \mathbb{R}^{m \times d}$, where m is the number of news items (articles) and d is the embedding dimensionality. Each row of the matrix corresponds to the embedding of a specific news article.

Table 1
Literature review summary.

Ref.	Model	Contributions	Limitations
[7]	Bimodal (CNN & LSTM)	captures sequential dependencies	limited to short textual data
[49]	Deep learning	enhanced accuracy.	high computational resources and overfitting
[30]	NLP & NER	i-domain and cross-domain analysis.	poor on text style in lengthy news
[50]	GAN & CNN	uses diverse data sources	Potential difficult standardisation
[51]	Text preprocessing	optimal preprocessing methods	weak analysis for entity relationships
[52]	BERT	improved sentence embeddings	poor validation in long dependencies
[53]	Pre-trained LM	high contextual understanding	restricted to educational texts
[54]	BERT models	effective on specific contexts	Turkish language only
[55]	BiLSTM	attention mechanisms	high computational requirements
[56]	ConvNet	operates on the web	specific for short web data
[6]	fine-tuned BERT	improved generalisation	overfitting with extensive fine-tuning
[57]	BERT-based	blended approach	complex, computationally demanding
[2]	BERT + RoBERTa	use of advanced models	high computational requirements
[26]	TF-IDF with MLP	analysed key linguistic patterns	reduced accuracy
[26]	BiLSTM CNN-LSTM	preprocessing pipeline	lacks interpretability

Our goal is to propose a model M that can be trained on the matrix $\mathbf{N}$ to learn a function $f : \mathbb{R}^d \rightarrow \{0,1\}$, where d is the embedding dimensionality and the model predicts the class $\hat{c}_i = M(\mathbf{N}_i) = f(\mathbf{N}_i)$ for each article i, with $\mathbf{N}_i$ representing the embedding of the ith article.

Word weights play an integral role in assessing the truthfulness of content, and we propose a new way to improve this process by integrating NER and RFC into a unified joint learning framework based on BERT.

3.1. The employed BERT model and its modification

Our Enhanced BERT model, based on the BERT_large architecture,[1] [58,59] is the core component in our joint learning framework for fake news detection. This model has been structurally modified from the traditional BERT to handle the contextual complexities of fake news by incorporating 24 transformer layers, each with 16 attention heads and 1024 hidden units. These architectural enhancements allow Enhanced BERT to capture both local and global dependencies in the input, enabling it to process diverse and contextually rich news content.

A distinguishing feature of our enhanced BERT implementation is the incorporation of task-specific attention masking directly within its multihead attention layers, eliminating the need for a separate Shared Task Presentation (STP) encoder as prior studies [60]. This approach allows the model to focus selectively on the portions of the input relevant to each task, such as named entities for NER, relational phrases for RFC, and coherence indicators for SD. By embedding task-specific masking directly into the BERT architecture, our model maintains a streamlined structure while remaining flexible enough to adapt to various tasks. This built-in task-specific masking optimises attention distribution across layers, enhancing task-specific focus without requiring additional encoding layers. This integrated approach preserves information clarity and maintains BERT's streamlined structure. By embedding task-specific masking directly, it optimises attention distribution, enhancing task-specific focus without extra encoding layers.

To further capture the complex features of fake news, we incorporate a hierarchical attention mechanism that operates at both the headline (sentence level) and the news body (document level). At the headline level, the model identifies key phrases or terms within the headline, capturing critical details that are often indicative of deceptive or misleading information. At the document level, attention focuses on broader narrative consistency across the news body, which is particularly valuable for stance detection. This broader scope enables the model to evaluate the alignment between a headline and the body of an article, helping to identify inconsistencies that may signal misinformation. This two-tiered attention mechanism – focusing on both headline-specific details and overall document coherence – enables Enhanced BERT to detect patterns of fake news at multiple levels of abstraction. The hierarchical attention can be represented as:

$$\text{Attention}^{\text{sentence}}_{\text{headline},i} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{headline}} \cdot \mathbf{K}^T_{\text{headline}}}{\sqrt{d_k}}\right) \tag{1}$$

$$\text{Attention}^{\text{document}}_{\text{body},j} = \text{softmax}\left(\frac{\mathbf{Q}_{\text{body}} \cdot \mathbf{K}^T_{\text{body}}}{\sqrt{d_k}}\right) \tag{2}$$

where Eq. (1) denotes the attention weight for the ith word in the headline, and Eq. (2) represents the attention weight for the jth word in the news body. Here, T is the total number of tokens in the sequence. This setup ensures that attention weights at the headline level focus on sentence-specific cues, while those at the body level capture the overall narrative.

The two-tiered attention mechanism enables the model to identify deceptive key phrases in headlines and maintain narrative consistency in the news body. This further improves Enhanced BERT's ability to detect fake news by analysing both linguistic details and overarching narrative structures, allowing it to identify fake news patterns at various levels of abstraction.

The self-attention mechanism in Enhanced BERT is further refined through a binary masking matrix $\mathbf{M}$, which controls the attention scores between words by dynamically adjusting their relevance for each task. The similarity between words i and j in the input is computed by the modified attention mechanism:

$$\text{Sim}(i,j) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + (\mathbf{M}_{i,j} - 1) \times C\right) \tag{3}$$

In this configuration, the constant C (a large negative value, e.g., -10^9) effectively excludes irrelevant word pairs from the attention mechanism by setting their scores close to zero when $\mathbf{M}_{i,j} = 0$, while allowing regular score computation when $\mathbf{M}_{i,j} = 1$. This selective attention mechanism enables Enhanced BERT to focus on contextually relevant text, enhancing interpretability and performance by prioritising named entities in NER tasks and aligning headlines with body content in stance detection tasks.

To mitigate the substantial memory demands associated with long-form content when using Enhanced BERT, we employ sequence handling techniques such as truncation and segment-level processing. These not only allow for better management of memory consumption while preserving model performance but also mitigate gradient explosions, which are likely to occur when using transformers to process extensive content.

Overall, Enhanced BERT's combination of task-specific attention masking, hierarchical attention, and refined self-attention mechanisms makes it uniquely capable of capturing both fine-grained details and broader contextual relationships within text. This setup allows the model to deliver context-sensitive embeddings that effectively support fake news detection, enabling it to differentiate between real and fake content with high accuracy across varied and complex news scenarios.

[1] https://www.kaggle.com/datasets/xhlulu/huggingface-bert

Table 2
Features of the FakeNewsNet dataset repository.

Features	PolitiFact		GossipCop	
	Fake	Real	Fake	Real
Total news articles	432	624	6,048	16,817
News articles with text content	353	400	785	16,765
News articles with social engagements	342	314	4,298	2,902
Articles with social engagements and news content	286	202	675	2,895
No. of tweets with replies	6,686	20,720	3,040	2,546
No. of tweets with likes	18,453	52,082	10,685	2,264
No. of tweets with retweets	13,226	42,059	7,614	5,025
Total no. of tweets	116,005	261,262	71,009	154,383

In the following sections, we refer to this model as our Enhanced BERT and describe our approach for constructing the task-specific masking matrix **M** in Section 4. This strategy enables Enhanced BERT to provide targeted attention, improving its effectiveness in fake news detection by leveraging contextually rich, task-specific embeddings.

3.2. Datasets

This study employs the FakeNewsNet data repository,[2] [61] including the PolitiFact dataset[3] and the GossipCop dataset,[4] both of which are acknowledged as standard benchmarks for the detection of fake news. PolitiFact contains actual and false statements by public and political figures in the U.S. Established in 2009, GossipCop enables actual news and false rumours in entertainment, focusing on celebrity gossip. Details on the datasets are displayed in Table 2.

To annotate the entities in the news articles, we used a pre-trained BERT-based NER model[5] [6], which is capable of identifying entities such as persons, organisations, and locations in text. This model was fine-tuned on our FakeNewsNet dataset to improve its performance on domain-specific entities, ensuring accurate identification of relevant entities within the context of news articles. For the RFC task, we leveraged several tools, including the spaCy-Transformers library, which integrates transformer-based models like BERT for relation extraction.[7] Similarly, the RFC model was fine-tuned on our dataset to classify relationships between identified entities, capturing the intricate relationships often present in fake news, such as exaggerated or misleading connections between entities. Details on the datasets are displayed in Table 2.

The evaluation of performance on our dataset is conducted using metrics such as Accuracy (α), the F1 score, precision π, recall (ρ), and the Area Under the (receiver operating characteristic) Curve (AUC).

4. The proposed joint learning framework

The purpose of JLFND is to advance the detection of fake news by processing text through multi-task learning with Enhanced BERT as a unified encoder. Unlike prior models that rely on STP-encoder, JLFND achieves task specialisation through task-specific attention masking within BERT's architecture, making it adaptable for NER, RFC, and SD tasks. Each task contributes a unique perspective: NER identifies key entities, RFC examines entity relationships, and stance detection checks for coherence between headlines and content. Our methodology leverages cross-task consistency and external knowledge validation, enabling robust, multi-dimensional fake news detection.

[2] https://github.com/KaiDMML/FakeNewsNet
[3] https://github.com/KaiDMML/FakeNewsNet/blob/master/dataset/politifact_fake.csv
[4] https://github.com/KaiDMML/FakeNewsNet/blob/master/dataset/gossipcop_fake.csv
[5] https://huggingface.co/dslim/bert-base-NER
[6] https://github.com/weizhepei/BERT-NER
[7] https://github.com/explosion/spacy-transformers

4.1. Architecture & general framework overview

The JLFND model in Fig. 1 leverages Enhanced BERT to provide context-rich embeddings that support a multi-task learning approach for the classification of fake news. By capturing both local and global dependencies, Enhanced BERT processes complex syntactic and semantic structures within a high-capacity 1024-dimensional hidden space. This structure allows JLFND to address three distinct yet interconnected tasks: NER, RFC, and SD, providing a comprehensive analysis framework to detect fake news.

During the fine-tuning process, the BERT model is trained on a dataset 3.2 consisting of text sequences from news articles, each paired with a true/false label indicating whether the article is real or fake. These text sequences are tokenised, and the BERT model generates rich contextual embeddings for each token. True/false labels play a crucial role in guiding model learning, as they are used in the final classification task to determine the authenticity of the news. Specifically, the labels directly influence the computation of the loss function, typically using cross-entropy loss, which is minimised during backpropagation. The embeddings generated by BERT are fed into separate classification heads for each task: NER, RFC, and SD. For NER, the embeddings help detect named entities, such as people or organisations. For RFC, embeddings are used to identify relationships between these entities, while for SD, embeddings are used to assess alignment and coherence between the headline and body of the article. Ultimately, the final classification—whether the article is real or fake—is predicted based on these task-specific embeddings, with the true/false labels playing a critical role in guiding the overall training process.

The Enhanced BERT model first tokenises the input text x into a sequence of tokens $\{w_1, w_2, \ldots, w_n\}$, where each token w_i is transformed into an embedding $h_i \in \mathbb{R}^d$. These embeddings pass through all 24 layers of Enhanced BERT, where each layer progressively refines the representations through self-attention and feed-forward transformations. This process generates a set of final contextualised embeddings $H_{\text{tokens}} = \{h_1^{\text{final}}, h_2^{\text{final}}, \ldots, h_n^{\text{final}}\}$, capturing nuanced semantic and syntactic information for each token. The embedding of the **[CLS]** token, denoted as $h_{\text{CLS}}^{\text{final}}$, serves as a summary representation of the entire input text, essential for the final classification decision.

Once BERT generates contextualised embeddings for the input sequence, these embeddings are utilised across the three task-specific heads: NER, RFC, SD. Each task-specific head applies a linear classifier to the corresponding token or sequence embeddings to produce predictions tailored to its objective. The NER head classifies tokens as entity types, the RFC head identifies and labels relationships between entities, and the SD head computes the semantic coherence between the headline and the body of the news article. These task-specific outputs are further aggregated to guide the final fake news classification, leveraging multi-task learning to enhance model robustness.

Algorithm 1 illustrates the training of the general framework. A detailed explanation of each step of the pseudocode is provided in the remainder of this section.

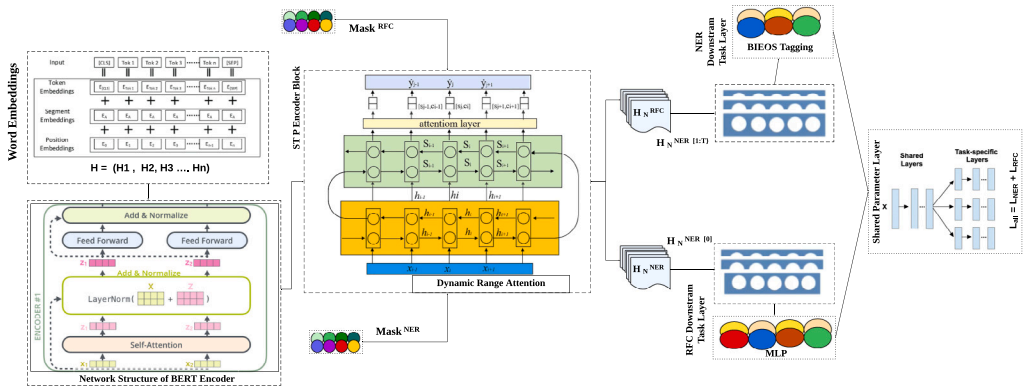


Fig. 1. Proposed model architecture diagram.

Algorithm 1 JLFND Pseudocode

Require: Dataset $N = \{N_1, N_2, \ldots, N_m\}$ and labels $C = \{c_1, c_2, \ldots, c_m\}$
Initialise Enhanced BERT ▷ Section 3.1
Tokenise dataset
Prepare NER tag ▷ BIEOS
for each epoch **do**
  for each sequence S_i in N **do**
    Tokenise sequence S_i with $[CLS]$ and $[SEP]$ tokens
    Encode S_i with Enhance BERT
    Prepare MASK_{NER} & MASK_{RFC} ▷ Section 4.3
    Obtain NER and RFC attention-focused embeddings
    Calculate NER output ▷ Section 4.4
    Calculate RFC output ▷ Section 4.5
    Calculate stance detection score ▷ Section 4.6
    Compute joint loss ▷ Section 4.7
    Backpropagate the gradients from the joint loss with respect to all model parameters ▷ Update the Enhanced BERT model parameters, task-specific attention masks, and task classifiers
  end for
end for
Use this trained model to predict label $\hat{c}$ for unseen sequences

4.2. Multi-task learning heads

The proposed multi-task framework allows simultaneous learning across tasks, improving overall model robustness while ensuring coherence between the outputs.

4.3. Attention mechanism and task-specific mask matrices

In place of an STP encoder, JLFND uses task-specific mask matrices within Enhanced BERT's attention mechanism. These masks are designed to selectively focus attention on relevant tokens for each task, such as entity-related tokens for NER and relational phrases for RFC. The mask value $\mathbf{M}_{i,j}^{\text{task}}$ for each token pair (i, j) is determined dynamically based on the task at hand. For example, in NER, the mask matrix is updated to assign a value of 1 to token pairs that correspond to named entities, and 0 to pairs that are unrelated to entities. In RFC, the mask matrix highlights pairs of tokens that represent potential relationships between entities, such as those involving verbs or prepositions connecting entities. In SD, the mask focuses on token pairs that connect the headline and body of the news article. Irrelevant token pairs receive a mask value of 0.

During the attention calculation, if the mask value $\mathbf{M}_{i,j}^{\text{task}}$ is 1, the token pair (i, j) is attended to normally. However, if the mask value is 0, the attention score between these tokens is suppressed by adding a large negative constant C (e.g., -10^9) to the attention computation. This prevents the attention mechanism from focusing on irrelevant token pairs, ensuring that only the most relevant relationships are considered during training.

Thus, the attention score for a given token pair is calculated as:

$$\text{Attention}_{i,j}^{\text{task}} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}} + (\mathbf{M}_{i,j}^{\text{task}} - 1) \times C\right) \tag{4}$$

This method ensures that the model dynamically adjusts its focus based on the task-specific relevance of token pairs, improving task-specific performance by focusing attention on the most important parts of the text.

4.4. The NER task

The NER task detects essential entities such as people, organisations, and locations in the text, which are crucial for the comprehension of the narrative. Token embeddings H_{tokens} feed a linear classifier that labels each token h_i^{final}, computing entity type probabilities (Eq. (5)), where $\mathbf{W}_{\text{NER}}$ and $\mathbf{b}_{\text{NER}}$ are learnable parameters.

$$P(z_i|h_i^{\text{final}}) = \text{softmax}(\mathbf{W}_{\text{NER}} h_i^{\text{final}} + \mathbf{b}_{\text{NER}}) \tag{5}$$

Here, z_i represents the predicted label for the ith token. The task minimises the cross-entropy loss function, where $y_{z,i}$ is the ground truth label for the ith token, indicating the actual entity class for that token.

$$L_{\text{NER}} = -\sum_{i=1}^{n} \sum_{z \in \mathcal{Z}} y_{z,i} \log P(z_i|h_i^{\text{final}}) \tag{6}$$

4.5. The RFC task

The RFC task identifies and classifies relationships between entities, capturing exaggerated or misleading connections common in fake news. For each pair of entities (e_i, e_j), a linear classifier predicts a relationship r from a set $\mathcal{R}$ of possible relationships, with the probability computed as shown in Eq. (7), where $\mathbf{W}_{\text{RFC}}$ and $\mathbf{b}_{\text{RFC}}$ are the classifier's parameters.

$$P(r|e_i, e_j) = \text{softmax}(\mathbf{W}_{\text{RFC}}[e_i; e_j] + \mathbf{b}_{\text{RFC}}) \tag{7}$$

In this equation, $y_{r,i,j}$ represents the ground truth label for the relationship between the entities e_i and e_j, where r denotes the specific relationship type. The cross-entropy loss function (Eq. (8)) is used for all pairs of entities.

$$L_{\text{RFC}} = -\sum_{(i,j)} \sum_{r \in \mathcal{R}} y_{r,i,j} \log P(r|e_i, e_j) \tag{8}$$

Table 3
Hyper-parameter values for Enhanced BERT.

Hyper-parameter	Value
Number of Hidden Layers	24
Dimensions of Encoder Layers	768
Learning Rate	Starts at 2e−5
Batch Size	32
Anneal Factor	5
Patience	5
Feed-forward Layers Dimensions in Encoder	3072
Number of Epochs	40
Dropout Rate	Variable, up to 0.2
Optimiser	AdamW
Weight Decay	0.005
Max Sequence Length	128
Training Duration	Extended duration to accommodate 40 epochs

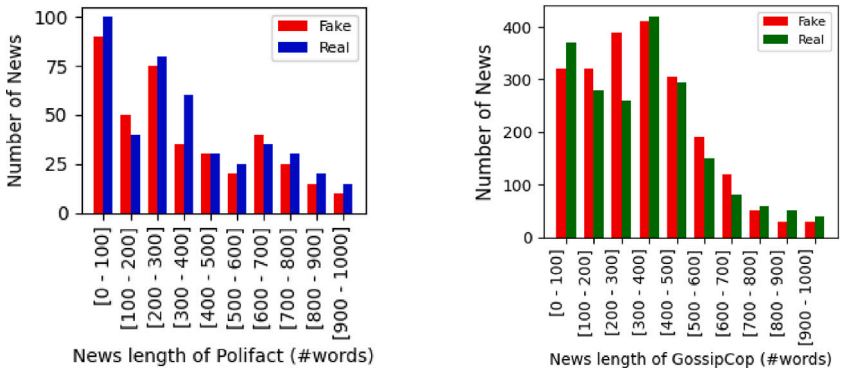


Fig. 2. Words frequency and news length.

4.6. The SD task

The SD task assesses the coherence between a headline and the body of the news content, identifying potential discrepancies. The stance score s is computed as the cosine similarity between the headline embedding h_{headline} and body embedding H_{body} projection matrices. For this task, we employ the stance loss function in Eq. (9) to enhance alignment for matching headline-body pairs and penalise discrepancies.

$$L_{\text{stance}} = \sum_{\text{pairs}} \max(0, 1 - y \cdot s) \tag{9}$$

4.7. Combined loss function

The combined loss function L_{final} integrates task-specific losses for NER (Eq. (6)), RFC (Eq. (8)), and SD (Eq. (9)) along with cross-task consistency (Eq. (10)) and knowledge (Eq. (11)) loss functions.

$$L_{\text{consistency}} = \sum_{i,j} \left| P(z_i | h_i^{\text{final}}) - P(r | e_i, e_j) \right| \tag{10}$$

$L_{\text{consistency}}$ ensures NER and RFC outputs remain consistent by aligning entity types and relationships, penalising any discrepancies.

$$L_{\text{knowledge}} = \sum_{i,j} \left| P(r | e_i, e_j) - P_{\text{fact}}(r | e_i, e_j) \right| \tag{11}$$

$L_{\text{knowledge}}$ ensures model predictions match external factual knowledge, avoiding contradictions. $P_{\text{fact}}(r|e_i, e_j)$ refers to the probability distribution of an external knowledge base.

The combined loss function L_{final} is computed as the weighted sum of each component (Eq. (12)).

$$L_{\text{final}} = \alpha L_{\text{NER}} + \beta L_{\text{RFC}} + \gamma L_{\text{stance}} + \delta L_{\text{consistency}} + \epsilon L_{\text{knowledge}} \tag{12}$$

The weights α, β, γ, δ, and ϵ are adjusted through cross-validation to balance the influence of each task and maintain consistency.

The task-specific mask matrices are dynamically adjusted during the fine-tuning process to ensure that the relationship between entities, their corresponding types, and the headline-body coherence are effectively learned and integrated.

5. Experimental setup

For the sake of replicability, we report the experimental and algorithm settings, as well as some implementation details.

The framework is implemented in Python using the PyTorch package.

To increase performance, we fine-tuned the model parameters using grid search and obtained the hyperparameter settings reported in Table 3. The NER and RFC tasks run simultaneously with randomly initialised downstream layers to start training. Note that the learning rate is initially 2×10^{-5} and then decreases linearly over 40 epochs with a decay rate of 4.75×10^{-7} per epoch. By the end of the 40th epoch, the learning rate will reach 1×10^{-6}, which ensures a smooth reduction in the learning rate throughout the training process. Additionally, the anneal factor is set to 5, meaning the learning rate is reduced by a factor of 5 during training. The patience parameter is set to 5 epochs, indicating that if there is no improvement in validation loss for 5 consecutive epochs, the learning rate will be reduced. To prevent overfitting, we employed a dropout rate of up to 0.2, where up to 20% of the units are dropped during training for regularisation.

We cleaned the data with NLTK,[8] and prepared it following a standard process in [62] which lowercases the text and removes stop words, symbols, null values, and any unknown characters before tokenisation. To address the class imbalance, we applied data balancing techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) [63], to ensure a well-distributed dataset across classes. Fig. 2 visually shows

[8] https://www.nltk.org/

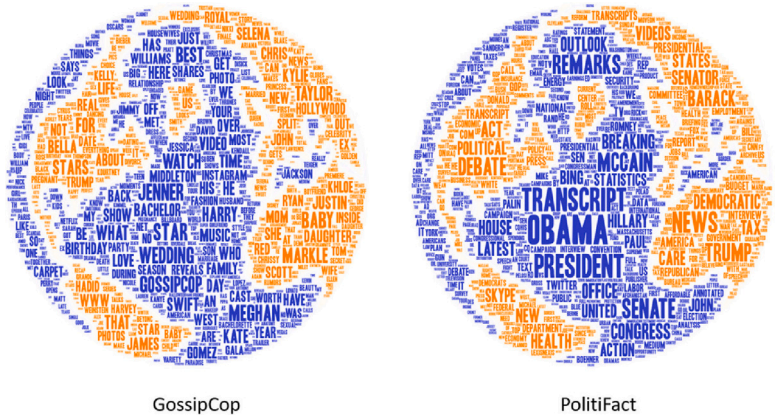
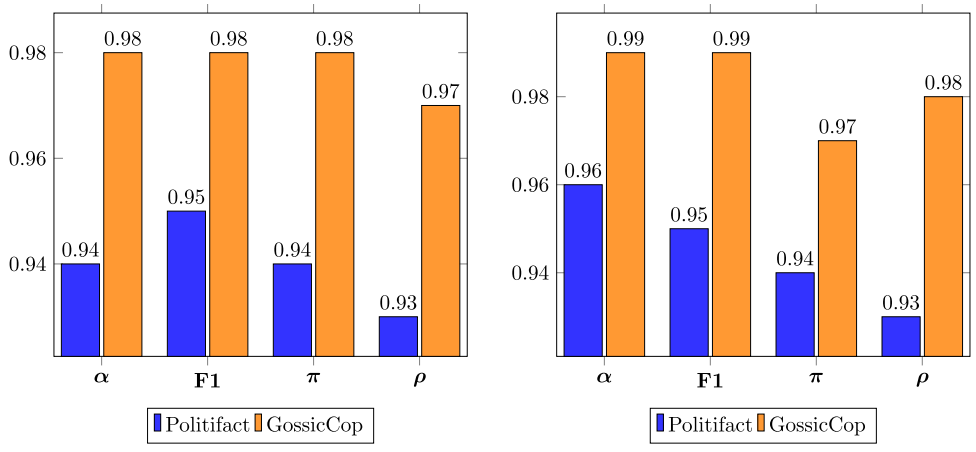


Fig. 3. Datasets Wordclouds.

Table 4
Evaluation metrics for proposed model on Politifact and GossipCop Datasets.

Dataset	Fake News				Real News			
	α	F1	π	ρ	α	F1	π	ρ
Politifact	0.94	0.95	0.94	0.93	0.96	0.95	0.94	0.93
GossipCop	0.98	0.98	0.98	0.97	0.99	0.99	0.97	0.98



(a) Fake News (b) Real News

Fig. 4. Evaluation metrics across the two datasets.

text frequencies by label, providing insights into data characteristics, including news length across both datasets. Each dataset is divided into training, testing, and validation subsets, following a 70%, 15%, and 15% distribution.

Fig. 3 presents the word clouds for each dataset, highlighting the most common terms in real and fake news samples.

Training and experiments are performed on a cloud server using a single A100 PCIe GPU.

6. Results and discussion

Numerical results show good performance for JLFND in handling long texts and capturing long-range dependencies on both datasets (Table 4).

In the GossipCop dataset, the model achieved superior performance in all metrics evaluated, with a notably high accuracy and F1 score of 0.99 in the detection of real news. We highlight this visually in Fig. 4.

We analyse the average distance from components to the centroid of a news sample to differentiate genuine information from falsehoods in large-scale text data. This is visualised using Kernel Density Estimate (KDE) plots in Fig. 5, showing distance distributions for training and testing splits of the GossipCop and Politifact datasets.

We were expecting slightly smaller distances for authentic news and more considerable distances for fake news. Interestingly, the results showed a significant difference in the mean distances for real and fake news: GossipCop with 0.476 and 1.39, and Politifact with 0.06 and 2.03. The KDE plots show narrow distributions, indicating low variance within each class and minimal overlap between real and fake news, highlighting significant class separation. This consistent differentiation across datasets and training-validation splits demonstrates the method's effectiveness in identifying fake news in long text sequences.

Fig. 6 shows a receiver operating characteristic curve with AUCs of 0.93 and 0.97 for Politifact and GossipCop, respectively, indicating near-perfect detection. The lower AUC for Politifact implies a slightly higher misclassification rate than GossipCop. The confidence intervals confirm the reliability of the results. Politifact's AUC had a 0.95 confidence interval of 0.91, 0.95, while GossipCop's AUC of 0.97 had an interval of 0.96, 0.98. The narrow intervals suggest reliable AUC values

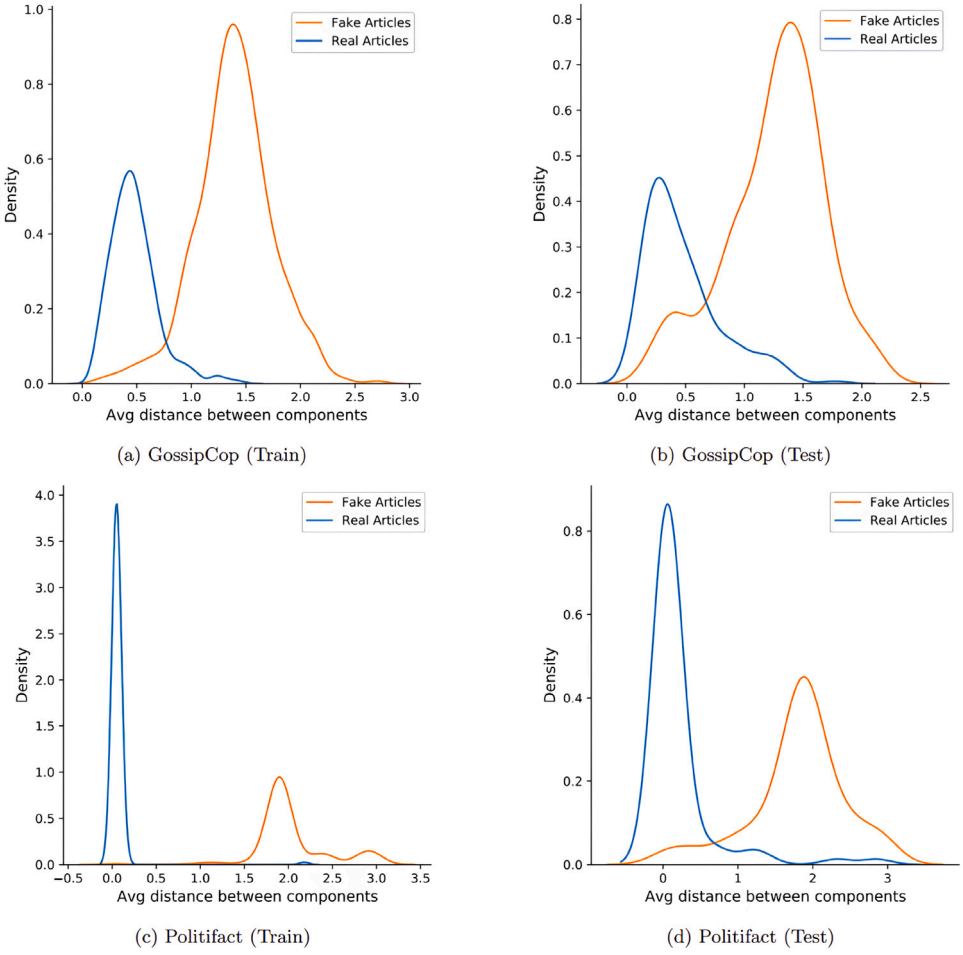


Fig. 5. Assessing modality discordance score on training and testing sets of Politifact and GossipCop.

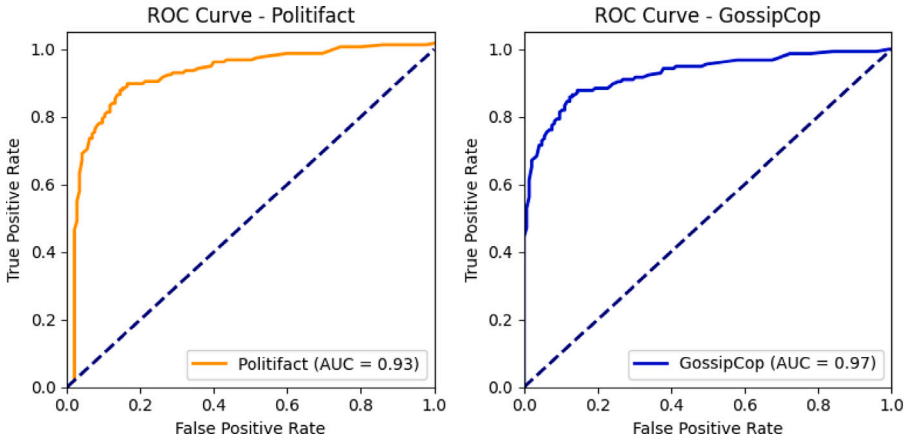


Fig. 6. The receiver operation curve of JLFND across the two datasets.

with minimal variability, reinforcing confidence in the model's ability to generalise to unseen data.

Fig. 7 indicates that the proposed model performs well in both datasets with high training and validation accuracy and low loss values. The proposed model fits the training data well and achieves a balance between training and validation accuracy, indicating it has learned the underlying patterns without overfitting. This is a result of several factors, including our fine-tuning procedure of the hyperparameters. The steady convergence of training and validation loss for both datasets shows that the model is not overfitting and has high generalisability.

6.1. Impact of NER and RFC

Table 5 shows the significant impact of NER and RFC on model performance.

Table 5 demonstrates the significant impact of NER and RFC on model performance. To obtain reliable estimates of model performance and standard deviation (std), we used a **70/15/15 dataset split**, allocating 70% for training, 15% for validation, and 15% for testing. We calculated std based on the results from multiple independent

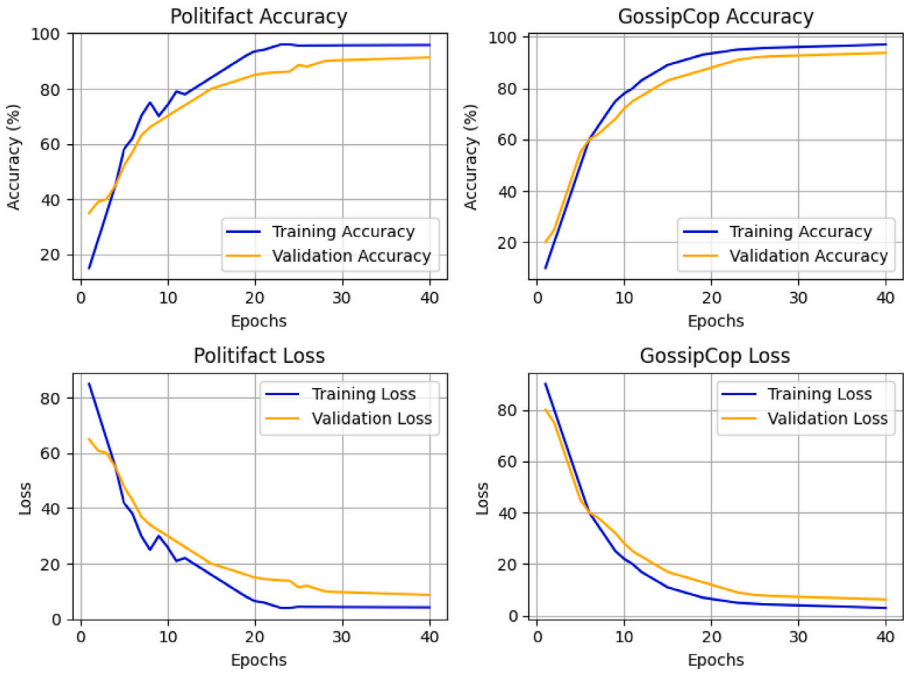


Fig. 7. Model performance in loss and accuracy.

Table 5
Performance analysis of NER and RFC.

Dataset	JLFND	JLFND$_{-NER}$	JLFND$_{-RFC}$	JLFND$_{-NER-RFC}$
Politifact	0.91 ± 0.05	0.82 ± 0.03	0.84 ± 0.05	0.76 ± 0.04
GossipCop	0.97 ± 0.05	0.83 ± 0.03	0.84 ± 0.01	0.81 ± 0.06

runs, ensuring robust evaluation by averaging outcomes across different model initialisation.

Including both NER and RFC (i.e., our JLFND model) yields the highest F1 scores (Politifact: 0.91 ± 0.05, GossipCop: 0.97 ± 0.05), highlighting the combined effectiveness of these components. Excluding NER results in a noticeable decrease in performance, underscoring its essential role in contextual understanding. Removing RFC also reduces scores, indicating its importance, though less critical than NER. The removal of both components significantly degrades performance, affirming their crucial role and showing that our method is well-designed with no redundant components.

Table 6 shows the impact of increasing batch sizes on the performance of the JLFND model in terms of F1 scores and processing times. A batch size of 32 is optimal, achieving maximum F1 scores of 0.94 for Politifact and 0.98 for GossipCop. Increasing the batch size to 64 only marginally improves the F1 score for GossipCop while significantly increasing the computation times in both datasets.

We also analysed the impact of varying sentence lengths and noted fluctuations in performance across sentences comprised of 100, 300, and 500 words. The majority of scholarly articles use a sentence length of 100 words for Politifact and a maximum of 300 words for GossipCop. However, our models demonstrate an ability to effectively manage longer statements with considerable efficiency. As illustrated in Table 7, the JLFND model exhibits enhanced performance with a sentence length of 200 words for Politifact and maintains consistent performance with 300 words for GossipCop, achieving an F1 Score of 0.86. Processing times, specifically 23 s for Politifact at 200 words and 21 s for GossipCop at 300 words, exhibit stability across different sentence lengths, highlighting the model's efficiency and scalability.

6.2. Cross dataset performance

To assess model generalisation, we conducted a cross-dataset performance analysis as shown in Table 8.

When the model is trained on the Politifact dataset and evaluated on the GossipCop dataset, the model maintains a reasonable level of performance when applied to a dataset with different characteristics. When the model is trained on the GossipCop dataset and assessed on the Politifact dataset, it shows enhanced performance metrics. This implies that the model exhibits more effective generalisation when the characteristics of the training and testing datasets are closely aligned.

The model exhibits robustness and adaptability, performing best when training and testing data align closely.

6.3. Impact of input size on model performance

We assessed JLFND across input sizes, namely 128, 256, and 512 tokens using BERT Base to evaluate effectiveness at lower dimensions. Smaller input sizes, such as 128 tokens, restrict the model's ability to capture longer contexts, whereas larger ones, like 256 or 512 tokens, enable the model to incorporate more contextual information and potentially enhance performance.

Our results, reported in Table 9, show that increasing the input size to 256 tokens improves the accuracy of the Politifact dataset. At 512 tokens, the model's accuracy and precision still increase, but there is a significant drop in terms of recall, resulting in a worse F1. Similarly, for the GossipCop dataset.

The Enhanced BERT model with a 768-token input performed excellently, achieving 0.94 accuracy for fake news and 0.96 for real news on the Politifact dataset, and 0.98 and 0.99 on the GossipCop dataset, respectively. Larger input sizes generally lead to better accuracy and AUC, but there is a precision–recall trade-off. The 512-token model achieved high precision but lower recall, indicating caution in identifying fake news, while the 128-token model had higher recall but lower precision, showing an aggressive approach. The 256-token model balanced precision and recall, offering a good compromise between context length and efficiency.

Table 6
JLFND evaluation at increasing batch sizes σ. Computational time τ is reported in seconds.

Dataset	$\sigma = 16$		$\sigma = 32$		$\sigma = 64$	
	F1	τ	F1	τ	F1	τ
Politifact	0.83	29	0.94	34	0.95	53
GossipCop	0.91	20	0.98	25	0.96	50

Table 7
Performances at increasing sentence lengths λ. Computational time τ is reported in seconds.

Dataset	$\lambda = 100$		$\lambda = 200$		$\lambda = 300$	
	F1	τ	F1	τ	F1	τ
Politifact	0.86	21	0.88	23	0.83	22
GossipCop	0.89	23	0.86	24	0.86	21

Table 8
Cross-Dataset analysis results.

Dataset		Metrics			
train	test	α	F1	π	ρ
Politifact	GossipsCop	82.12	83.24	83.56	83.79
GossipsCop	Politifact	88.97	89.10	89.19	89.15

Table 9
Evaluation at increasing input sizes.

Dataset	Input size	α	π	ρ	F1	AUC
Politifact	128	0.741	0.486	0.797	0.602	0.860
	256	0.821	0.630	0.708	0.674	0.872
	512	0.852	0.856	0.631	0.763	0.882
	768	0.940	0.940	0.930	0.950	0.930
GossipCop	128	0.835	0.574	0.770	0.628	0.851
	256	0.884	0.638	0.723	0.743	0.863
	512	0.912	0.914	0.614	0.862	0.881
	768	0.980	0.980	0.970	0.980	0.970

6.4. Ablation study

We report the results of the ablation analysis in Table 10 to evaluate the impact of individual modules in diverse configurations of the JLFND framework. These results are obtained by replacing our Enhanced BERT module with alternative NLP models, including RoBERTa [64] (referred to as $\text{JLFND}_{\text{RoBERTa}}$), BiLSTM [65] (referred to as $\text{JLFND}_{\text{BiLSTM}}$), Text CNN [66] (as $\text{JLFND}_{\text{TextCNN}}$), Text RNN [67] (referred to as $\text{JLFND}_{\text{TextRNN}}$), and standard BERT [24] (referred to as $\text{JLFND}_{\text{BERT}}$). While TextCNN is traditionally employed for text classification tasks, it has been included in our ablation study to evaluate its performance in a multi-task setting within the JLFND framework. Although TextCNN does not directly generate embeddings like BERT-based models, it creates token-level representations used in classification. We assess how this model performs for tasks like NER, RFC, and SD within JLFND.

We also examine JLFND when it is deprived of one of its modules (indicated with a minus sign '–' in Table 10) and the performances of our Enhance BERT on its own and with the addition of relevant modules (indicated with a plus sign '+ ' in Table 10).

Our analysis highlights that the full JLFND model, which integrates Task-Specific Masking, Hierarchical Attention (both sentence-level and document-level), Multi-Task Learning, and Consistency Losses, always achieves the highest performances. In the absence of Task-Specific Masking, we observe a reduction in NER and RFC performance, as the model cannot focus as effectively on task-relevant tokens. Removing Hierarchical Attention, particularly the document and sentence levels, deteriorates performances. Without Multi-Task Learning, tasks (NER, RFC, SD) function separately, diminishing effectiveness by ignoring inter-task dependencies. Omitting Consistency Losses decreases coherence and factual alignment across tasks, particularly on the Politifact dataset. The alternative models reveal the strength of our Enhanced BERT in the JLFND framework. Standard BERT outperforms Text CNN and Text RNN in terms of contextual understanding, while RoBERTa demonstrates robust performance, reflecting its architectural optimisations for NLP [64]. However, using Enhanced BERT yields the highest classification effectiveness across both datasets, underscoring the value of its tailored components in handling complex relationships and nuanced semantic structures in fake news detection.

6.5. Comparative analysis

We comprehensively compare baseline algorithms from Section 2, including deep neural networks and state-of-the-art transformer models, to JLFND. Results are reported in Table 11.

6.6. Limitations and future improvements

While JLFND marks a substantial advancement in fake news detection, there are several areas for potential improvement. Integrating external databases could enhance the model's accuracy by cross-referencing content with factual sources, particularly benefiting cases that require complex, context-dependent analysis. Additionally, incorporating interpretability techniques, such as SHAP values or attention heatmaps, would make the model's decision-making process more transparent, which is essential for ensuring trust and reliability, especially in sensitive applications.

7. Conclusion

We introduce the JLFND framework and demonstrate that our methodology outperforms its competitors in two widely used FakeNewsNet data sets: Politifact and GossipCop. Our model achieves state-of-the-art results, with accuracies of 0.94 and 0.98 on these datasets, respectively.

Our approach is novel and incorporates several unique features:

Table 10
Ablation study results on Politifact and GossipCop datasets.

Model	Politifact		GossipCop	
	α	F1	α	F1
JLFND	0.94	0.95	0.98	0.98
JLFND – Task-Specific Masking	0.91	0.92	0.95	0.96
JLFND – Hierarchical Attention	0.90	0.91	0.94	0.95
JLFND – Sentence-Level Attention	0.92	0.93	0.96	0.96
JLFND – Document-Level Attention	0.91	0.92	0.95	0.95
JLFND – Multi-Task Learning	0.89	0.90	0.93	0.93
JLFND – Consistency Losses	0.92	0.93	0.96	0.97
$JLFND_{TextCNN}$	0.86	0.87	0.88	0.89
$JLFND_{TextRNN}$	0.89	0.89	0.90	0.91
$JLFND_{BiLSTM}$	0.86	0.86	0.88	0.87
$JLFND_{BiLSTM+Att.}$	0.86	0.87	0.89	0.90
$JLFND_{BERT}$	0.88	0.89	0.91	0.91
$JLFND_{RoBERTa}$	0.90	0.91	0.93	0.94
Enhanced BERT	0.91	0.92	0.95	0.96
Enhanced BERT + RFC	0.92	0.92	0.95	0.96
Enhanced BERT + Mask_NER	0.93	0.92	0.97	0.97
Enhanced BERT + Joint learning	0.94	0.95	0.98	0.98

Table 11
Comparative analysis results (best result per metric per dataset in boldface).

Ref.	Year	Model	Politifact				GossipCop			
			α	F1	π	ρ	α	F1	π	ρ
[46]	2020	SAFE	0.87	0.88	0.90	0.89	0.83	0.85	0.93	0.89
[11]	2022	XG Boost	0.92	0.91	0.91	0.92	0.91	0.91	0.91	0.90
[26]	2024	BERT	0.91	0.92	0.93	0.91	0.85	0.82	0.76	0.89
[42]	2024	Small-BERT	0.78	0.80	0.78	0.77	0.70	0.72	0.72	0.71
[45]	2024	GCNs-MT	0.85	0.84	0.80	0.92	0.97	0.97	0.97	0.96
[43]	2024	L-Agent	0.88	0.88	0.89	0.88	0.83	0.83	0.83	0.83
[14]	2024	GAT-GANM	0.83	0.85	0.85	0.86	0.95	0.96	0.95	0.95
[44]	2024	KEAP	0.92	0.91	0.91	0.92	0.91	0.91	0.91	0.90
[48]	2024	GraphSage	0.89	0.89	0.92	0.85	0.77	0.77	0.82	0.70
[26]	2024	PA Model	0.83	0.82	0.83	0.81	0.80	0.72	0.72	0.72
[47]	2024	MRE-FND	0.92	0.91	0.92	0.91	0.83	0.81	0.83	0.81
—		JLFND	**0.94**	**0.95**	**0.94**	**0.93**	**0.98**	**0.98**	**0.98**	**0.97**

- We use **hierarchical attention mechanisms** that capture both sentence-level (headline) and document-level (body) context, which is crucial for assessing the coherence between headline and content.
- A **joint learning framework** integrates RFC, NER, and SD tasks with the Enhanced BERT model. This setup allows each task to benefit from shared parameters and multitask learning, improving model generalisation.
- **Task-specific masking matrices** are used to refine attention distribution, enhancing focus on relevant tokens for each task. This is particularly effective for RFC and NER task performance.
- We use **cross-task consistency losses** to ensure coherence and alignment across tasks, enhancing the model's ability to maintain factual consistency.
- A last contribution is made by analysing the **average distance of components to the centroid** of a news sample, effectively distinguishing genuine information from falsehoods in large-scale text data.

Our findings suggest that JLFND is not only effective in fake news detection but also shows promise for applications in other areas involving the processing of long texts, such as automatic text summarisation.

Currently, JLFND has not been tested in a multi-class fake news classification setting, which would involve distinguishing subtle differences between classes through more advanced analysis. We aim to address this challenge in future work.

CRediT authorship contribution statement

Muhammad Abdullah: Writing – original draft, Visualization, Investigation, Validation, Methodology, Software, Formal analysis. **Zan Hongying:** Supervision, Conceptualization. **Arifa Javed:** Writing – original draft, Conceptualization. **Orken Mamyrbayev:** Conceptualization. **Fabio Caraffini:** Writing – review & editing, Visualization, Validation. **Hassan Eshkiki:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fabio Caraffini reports article publishing charges was provided by Swansea University. If there are other authors, they 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

Data are publicly available and referenced in the article.

References

[1] F. Olan, U. Jayawickrama, E.O. Arakpogun, J. Suklan, S. Liu, Fake news on social media: the impact on society, Inf. Syst. Front. (2022) 1–16.
[2] L.B. Angizeh, M.R. Keyvanpour, Detecting fake news using advanced language models: BERT and roberta, in: 2024 10th International Conference on Web Research, ICWR, IEEE, 2024, pp. 46–52.
[3] S. Biradar, S. Saumya, A. Chauhan, Combating the infodemic: COVID-19 induced fake news recognition in social media networks, Complex Intell. Syst. 9 (3) (2023) 2879–2891.
[4] S. Nieminen, V. Sankari, Checking politifact's fact-checks, Journal. Stud. 22 (3) (2021) 358–378.

[5] A. Areshey, H. Mathkour, Transfer learning for sentiment classification using bidirectional encoder representations from transformers (BERT) model, Sensors 23 (11) (2023) 5232.

[6] S. Qin, M. Zhang, Boosting generalization of fine-tuning BERT for fake news detection, Inf. Process. Manage. 61 (4) (2024) 103745.

[7] A. Abdullah, M. Awan, M. Shehzad, M. Ashraf, Fake news classification bimodal using convolutional neural network and long short-term memory, Int. J. Emerg. Technol. Learn. 11 (2020) 209–212.

[8] Y. Bai, M. Chen, P. Zhou, T. Zhao, J. Lee, S. Kakade, H. Wang, C. Xiong, How important is the train-validation split in meta-learning? in: International Conference on Machine Learning, PMLR, 2021, pp. 543–553.

[9] J. Jouhar, A. Pratap, N. Tijo, M. Mony, Fake news detection using python and machine learning, Procedia Comput. Sci. 233 (2024) 763–771, http://dx.doi.org/10.1016/j.procs.2024.03.265, URL: https://www.sciencedirect.com/science/article/pii/S1877050924006252. 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024).

[10] M. Iftikhar, A. Ali, Fake news detection using machine learning, in: 2023 3rd International Conference on Artificial Intelligence, ICAI, IEEE, 2023, pp. 103–108.

[11] K.A. Qureshi, R.A.S. Malick, M. Sabih, H. Cherifi, Deception detection on social media: A source-based perspective, Know.- Based Syst. 256 (C) (2022) http://dx.doi.org/10.1016/j.knosys.2022.109649, URL: https://doi.org/10.1016/j.knosys.2022.109649.

[12] M.I. Nadeem, K. Ahmed, D. Li, Z. Zheng, H.K. Alkahtani, S.M. Mostafa, O. Mamyrbayev, H. Abdel Hameed, EFND: A semantic, visual, and socially augmented deep framework for extreme fake news detection, Sustainability 15 (1) (2022) 133.

[13] B. Hu, Z. Mao, Y. Zhang, An overview of fake news detection: From a new perspective, Fundam. Res. (2024).

[14] Q. Chang, X. Li, Z. Duan, Graph global attention network with memory: A deep learning approach for fake news detection, Neural Netw. 172 (2024) 106115.

[15] M.I. Nadeem, S.A.H. Mohsan, K. Ahmed, D. Li, Z. Zheng, M. Shafiq, F.K. Karim, S.M. Mostafa, Hyprobert: A fake news detection model based on deep hypercontext, Symmetry 15 (2) (2023) 296.

[16] M. Bilal, A.A. Almazroi, Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews, Electron. Commer. Res. 23 (4) (2023) 2737–2757.

[17] S. Islam, N. Ab Ghani, M. Ahmed, A review on recent advances in deep learning for sentiment analysis: Performances, challenges and limitations, Compusoft 9 (7) (2020) 3775–3783.

[18] W. Khan, A. Daud, K. Khan, S. Muhammad, R. Haq, Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends, Nat. Lang. Process. J. (2023) 100026.

[19] B. Xie, Q. Li, Detecting fake news by RNN-based gatekeeping behavior model on social networks, Expert Syst. Appl. 231 (2023) 120716.

[20] M. Mende, V.O. Ubal, M. Cozac, B. Vallen, C. Berry, Fighting infodemics: Labels as antidotes to mis-and disinformation?! J. Public Policy Mark. 43 (1) (2024).

[21] M.I. Nadeem, K. Ahmed, D. Li, Z. Zheng, H. Naheed, A.Y. Muaad, A. Alqarafi, H. Abdel Hameed, SHO-CNN: a metaheuristic optimization of a convolutional neural network for multi-label news classification, Electronics 12 (1) (2022) 113.

[22] S. Raza, C. Ding, Fake news detection based on news content and social contexts: a transformer-based approach, Int. J. Data Sci. Anal. 13 (4) (2022) 335–362.

[23] L. Peng, S. Jian, Z. Kan, L. Qiao, D. Li, Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection, Inf. Process. Manage. 61 (1) (2024) 103564.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[26] L. Aslan, M. Ptaszynski, J. Jauhiainen, Are Strong Baselines Enough? False News Detection with Machine Learning, Preprints, 2024.

[27] M. Sathvik, M.K. Mishra, S. Padhy, Fake news detection by fine tuning of bidirectional encoder representations from transformers, Authorea Prepr. (2023).

[28] C. Buntain, J. Golbeck, Automatically identifying fake news in popular Twitter threads, in: 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 208–215, http://dx.doi.org/10.1109/SmartCloud.2017.40.

[29] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explor. Newsl. 19 (1) (2017) 22–36.

[30] C.-M. Tsai, Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis, Electronics 12 (17) (2023) 3676.

[31] V. Sanh, Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter, 2019, arXiv preprint arXiv:1910.01108.

[32] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021) 1–74.

[33] Q. Yu, Z. Wang, K. Jiang, Research on text classification based on bert-bigru model, in: Journal of Physics: Conference Series, vol. 1746, (1) IOP Publishing, 2021, 012019.

[34] Z. Sun, G. Zhao, R. Scherer, W. Wei, M. Woźniak, Overview of capsule neural networks, J. Inter. Technol. 23 (1) (2022) 33–44.

[35] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, Y. Yu, Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1212–1220, http://dx.doi.org/10.1145/3474085.3481548, URL: https://doi.org/10.1145/3474085.3481548.

[36] J. Chen, C. Jia, H. Zheng, R. Chen, C. Fu, Is multi-modal necessarily better? Robustness evaluation of multi-modal fake news detection, IEEE Trans. Netw. Sci. Eng. 10 (6) (2023) 3144–3158, http://dx.doi.org/10.1109/TNSE.2023.3249290.

[37] J. Wang, S. Qian, J. Hu, R. Hong, Positive unlabeled fake news detection via multi-modal masked transformer network, IEEE Trans. Multimed. 26 (2023) 234–244.

[38] S. Qian, J. Wang, J. Hu, Q. Fang, C. Xu, Hierarchical multi-modal contextual attention network for fake news detection, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 153–162.

[39] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2560–2569.

[40] D. Khattar, J.S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.

[41] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2018, pp. 849–857.

[42] F. Folino, G. Folino, M. Guarascio, L. Pontieri, P. Zicari, Towards data-and compute-efficient fake-news detection: An approach combining active learning and pre-trained language models, SN Comput. Sci. 5 (5) (2024) 470.

[43] X. Li, Y. Zhang, E.C. Malthouse, Large language model agent for fake news detection, 2024, arXiv preprint arXiv:2405.01593.

[44] X. Che, G. Yang, Y. Chen, Q. Li, Detecting fake information with knowledge-enhanced AutoPrompt, Neural Comput. Appl. 36 (14) (2024) 7725–7742.

[45] Q. Chang, X. Li, Z. Duan, A novel approach for rumor detection in social platforms: Memory-augmented transformer with graph convolutional networks, Knowl.-Based Syst. 292 (2024) 111625, http://dx.doi.org/10.1016/j.knosys.2024.111625, URL: https://www.sciencedirect.com/science/article/pii/S0950705124002600.

[46] X. Zhou, J. Wu, R. Zafarani, SAFE: similarity-aware multi-modal fake news detection, 2020, CoRR abs/2003.04981. URL: https://arxiv.org/abs/2003.04981. arXiv:2003.04981.

[47] L. Han, X. Zhang, Z. Zhou, Y. Liu, A multifaceted reasoning network for explainable fake news detection, Inf. Process. Manage. 61 (6) (2024) 103822, http://dx.doi.org/10.1016/j.ipm.2024.103822, URL: https://www.sciencedirect.com/science/article/pii/S030645732400181X.

[48] P. Kapadia, An Explainable Approach to Multi-contextual Fake News Detection.

[49] M. Samadi, M. Mousavian, S. Momtazi, Deep contextualized text representation and learning for fake news detection, Inf. Process. Manage. 58 (6) (2021) 102723.

[50] S. Mishra, P. Shukla, R. Agarwal, Analyzing machine learning enabled fake news detection techniques for diversified datasets, Wirel. Commun. Mob. Comput. 2022 (2022) 1–18.

[51] C.P. Chai, Comparison of text preprocessing methods, Nat. Lang. Eng. 29 (3) (2023) 509–553.

[52] O. Galal, A.H. Abdel-Gawad, M. Farouk, Rethinking of BERT sentence embedding for text classification, 2024.

[53] Z. Liu, X. He, L. Liu, T. Liu, X. Zhai, Context matters: A strategy to pre-train language model for science education, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 666–674.

[54] G.K. Koru, Ç. Uluyol, Detection of turkish fake news from tweets with BERT models, IEEE Access (2024).

[55] M. Choudhary, S.S. Chouhan, E.S. Pilli, S.K. Vipparthi, BerConvoNet: A deep learning framework for fake news classification, Appl. Soft Comput. 110 (2021) 107614.

[56] D.K. Vishwakarma, P. Meel, A. Yadav, K. Singh, A framework of fake news detection on web platform using ConvNet, Soc. Netw. Anal. Min. 13 (1) (2023) 24.

[57] S. Mahadevan sr, S. Ahmad, BERT based blended approach for fake news detection, J. Big Data Artif. Intell. 2 (1) (2024).

[58] M.V. Koroteev, BERT: a review of applications in natural language processing and understanding, 2021, arXiv preprint arXiv:2103.11943.

[59] Y. Hao, L. Dong, F. Wei, K. Xu, Visualizing and understanding the effectiveness of BERT, 2019, arXiv preprint arXiv:1908.05620.

[60] W. Shishah, Fake news detection using BERT model with joint learning, Arab. J. Sci. Eng. 46 (9) (2021) 9115–9127.

[61] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (3) (2020) 171–188.

[62] V. Maslej-Krešňáková, M. Sarnovskỳ, P. Butka, K. Machová, Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification, Appl. Sci. 10 (23) (2020) 8631.

[63] A. Fernández, S. Garcia, F. Herrera, N.V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, J. Artificial Intelligence Res. 61 (2018) 863–905.

[64] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, 2019, URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[65] G. Liu, J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, Neurocomputing 337 (2019) 325–338, http://dx.doi.org/10.1016/j.neucom.2019.01.078, URL: https://www.sciencedirect.com/science/article/pii/S0925231219301067.

[66] I. Alshubaily, Textcnn with attention for text classification, 2021, arXiv preprint arXiv:2108.01921.

[67] F. Xia, et al., Label oriented hierarchical attention neural network for short text classification, Acad. J. Eng. Technol. Sci. 5 (1) (2022) 53–62.