L^{3}B – Lies Lie in Linguistic Behavior: Learning Verbal Indicators for Content Veracity Classification

Anonymous ACL submission

Abstract

Unverified content poses significant challenges by disrupting content veracity and integrity, thereby making effective content classification approaches crucial. Currently, content veracity classification methods primarily use supervised machine learning models, which, despite high accuracy, lack generalizability due to heavy reliance on raw content data. To address this issue, we propose a behavior-aware classification model (L^{3}B) leveraging latent linguistic behavior and external social context to extract contextually grounded features, reducing reliance on content data and sensitivity to data biases. First, we extract the verbal features from news content as linguistic behavior features and capture nuanced behavior indicators of content veracity. Then, a knowledge-based linking scheme is designed to incorporate social context, aligning extracted verbs with those derived from linked social context using semantic similarity. Finally, we feed the textual, behavioral, and contextual features into a Transformer-based classifier to fuse these features and then classify the content veracity (i.e., high or low veracity). Experimental results on public datasets demonstrate that our model outperforms most advanced classification approaches and has improved generalizability across diverse datasets, highlighting the effectiveness and robustness of our proposed model.

1 Introduction

Nowadays, with the rapid rise of social media platforms, such as X (Twitter), Instagram, and TikTok, an increasing number of individuals heavily rely on these online platforms for communication, information dissemination, and education, especially during the pandemic (Tsao et al., 2021). Though the conveniences brought by social media, the content veracity of information disseminated still falls short of media standards and social expectations, compared to traditional media platforms, e.g., television and newspapers (Shu et al., 2017; Zhou and Zafarani, 2020). A large volume of unverified or distorted content is easily produced and propagated through social media platforms (Ahmed et al., 2022), especially using artificial intelligence tools to fabricate news content (Zhou et al., 2023). Given that content veracity refers to the degree to which a news or article aligns with authenticity, it plays a critical role in maintaining content integrity, where low-veracity content (e.g., spam, rumor, false information, etc.) has significant negative impacts on individual and society, such as social trust, government authority, and information credibility (Thorson et al., 2010; Bhattarai et al., 2021; Mazzeo et al., 2021). Consequently, addressing low-veracity content propagation has become crucial in the areas of social media, mass communication, and public health. Technically, automatic models are developed to identify and classify the low-veracity content on social media platforms, thereby mitigating the negative impacts brought by low-veracity content (Guo et al., 2020; Yang et al., 2023; Shi et al., 2023).

While content veracity classification methods have achieved significant advancements, these methods still struggle with feature complexity, dataset biases, and generalizability issues across different application scenarios (Zubiaga et al., 2018; Abdali et al., 2024). High-quality annotated data is scarce (Bondielli and Marcelloni, 2019), leading to existing models compromising classification performance on unseen content data. To overcome these limitations, efficient, unbiased, and scalable classification frameworks for content veracity are needed, which are capable of adapting to new instances and social contexts.

To address these issues, in this paper, we integrate social content and context features for developing an efficient and scalable content veracity classification model (L^{3}B). By exploring these additional features, our model can reduce bias and reliance across different datasets. Specially, our contributions are summarized below:

- Firstly, we extract the verb features from news content as linguistic behavior features and capture nuanced behavior indicators of content veracity.
- Secondly, a knowledge-based linking scheme is designed to incorporate social context, further refining linguistic behavior features and mitigating classification bias and distribution shifts.
- Finally, we feed the text, linguistic behavior, and social context features into a transformer-based classifier to fuse these features and classify content veracity.
- Experimental results on public datasets demonstrate that our model outperforms most advanced classification approaches, highlighting the effectiveness and scalability of our proposed model.

2 Related Work

2.1 Content-based methods

Traditional classification methods focus on internal news content features and external fact-checking resources to detect content veracity (Vlachos and Riedel, 2014; Hassan et al., 2015; Guo et al., 2022). For instance, the fact-checking approaches can identify and classify the low-veracity content by using the external knowledge sources to fact-check the news content (Etzioni et al., 2008; Wu et al., 2014; Shi and Weninger, 2016; Vo and Lee, 2018). However, these fact-checking approaches are time-consuming and demand human annotations, limiting the scalability and efficiency in content veracity classification.

Today, machine learning (ML) and natural language processing (NLP) methods (Kadhim, 2019; Su et al., 2020) have emerged as advanced tools to classify news text into one or more predefined classes, such as true or false. Traditional ML methods, such as support vector machine, random forest, and decision tree, are commonly used in news content classification; however, these methods usually require hand-crafted features and struggle with complex text features, thus compromising performance (Minaee et al., 2021). Along with neural networks being boosted, deep learning frameworks have further enhanced the classification performance by extracting complex content features and capturing nuanced semantic features, such as convolutional neural networks (CNNs) (Kim, 2014; Wang, 2017; Ruchansky et al., 2017; Guo et al., 2019; Kaliyar et al., 2020), recurrent neural networks (RNNs) (Ruchansky et al., 2017; Ma et al., 2016), and long short-term memory (LSTM) (Sachan et al., 2019; Ma et al., 2020). Kaliyar et al. (Kaliyar et al., 2020) proposed a deep CNN model for binary classification (i.e., true or false) of news content compared to classical CNN and LSTM structures, where it explores pre-trained word embedding and multiple hidden layers to extract text features. In addition, attention networks integrated different features extracted from different latent aspects of news articles to improve classification accuracy (Yang et al., 2016; Mishra and Setty, 2019; Linmei et al., 2019; Sun and Lu, 2020; Yun et al., 2023; Kim and Hwang, 2024). For example, Yang et al. (Yang et al., 2016) proposed a hierarchical attention network (HAN) to capture the hierarchical structure of documents and employ the word-level and sentence-level attentions. Kim et al. (Kim and Hwang, 2024) employed attention mechanisms to identify semantically similar words within sentences and then augment these sentences using synonym replacements. Additionally, graph convolutional networks (GCNs) (Yao et al., 2018; Haider Rizvi et al., 2025) have been applied to textual content classification tasks, which construct document-level and corpus-level graphs to learn relationships among words, documents, and corpus.

With the aid of pre-trained knowledge embeddings, the transformer-based models have advanced the classification accuracy of low-veracity content in news articles (Liu et al., 2019; Croce et al., 2020; Kaliyar et al., 2021; Xiong et al., 2021; Van Nooten and Daelemans, 2025). Combining the bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) with a CNN structure, Kaliyar et al. (Kaliyar et al., 2021) proposed a BERT-based news classification model, where it inputs the BERT embeddings into one-dimensional CNN layers and thus classifies news documents using local features and global dependencies. Along with the data structure and modality extending, multimodal approaches are proposed to handle more intricate classification tasks for content veracity across text, image, video, audio data, or multiple languages (Conneau and Lample, 2019; Segura-Bedmar and Alonso-Bartolome, 2022; Abdali et al., 2024; Wu et al., 2024; Zeng et al., 2024). For example, Wu et al. (Wu et al., 2024) emphasized the substantive content over stylistic features, using Large Language Models (LLMs) to reframe news articles and focus on content veracity. Though LLMs emerged with powerful capability of processing multimodal features, LLMs still require a large volume of data to update the known knowledge and maintain performance and reliability.

2.2 Context-based methods

For further exploiting the source and content features to classify content veracity, the credibility-based methods (Popat, 2017; Zhang et al., 2018; Deng et al., 2025) were proposed, which could extract the source and content credibility features to identify high-veracity news from unreliable ones, thereby enhancing model performance. To explore the user behavior, engagements, and interactions on social media, the social relationship-aware approaches (Shu et al., 2019; Ghenai and Mejova, 2018; Shu et al., 2020; Dou et al., 2021; Teng et al., 2022; Su et al., 2023) were proposed, which can capture the users' relationships, news content, and dissemination patterns to improve classification accuracy. For instance, Shu et al. (Shu et al., 2019) presented a tri-relationship-based veracity classification framework of false news content (TriFN), where TriFN explores the tri-relationship among publishers, news pieces, and users to differentiate false and true articles. Zhang et al. (Zhang et al., 2024) explored the heterogeneous subgraph transformer (HeteroSGT) to classify articles via the heterogeneous graph by unearthing the relationships among news topics, entities, and content.

To understand the propagation patterns of low-veracity content within social networks, the network-based methods (Zhou and Zafarani, 2019) were suggested, where these methods focus on the interactions among spreaders and their influence on information propagation. Ma et al. (Ma et al., 2018) presented tree-structured recursive neural networks to model the propagation pattern of tweets for detecting rumors on social media. Typically, graph-based approaches were proposed (Bian et al., 2020; Fu et al., 2022) to explore the potential of graph structure in modeling social context structures, including knowledge-driven (Wang et al., 2018; Dun et al., 2021), propagation-based (Zhu et al., 2024), and context-aware approaches (Shang et al., 2024; Li et al., 2025). For instance, the propagation-based models (Zhu et al., 2024) focus on the dynamics of information dissemination within social networks, therefore identifying content veracity based on dissemination patterns.

3 Methodology

In this section, we first introduce the fundamental framework of our proposed L^{3}B model, as shown in Figure 1. Next, the verb-extraction module will be presented for extracting verbs from news content and then deriving the verb-based linguistic behavior features. Then, we adopt a knowledge-based linking scheme to incorporate social context features for refining linguistic behavior features, based on the similarity between the extracted verbs and social context verbs. Finally, the combined features, including content features, behavior features, and context features, are input into a transformer-based classifier to classify content veracity, where these features will be fused and then fed into the final layer.

3.1 Definitions

News articles usually involve most of the practical elements, such as sentiment, behavior, interaction, etc. Generally, these elements will be represented by nouns, verbs, adjectives, etc., or hidden behind the words and sentences in the article content. In addition, these elements will cover rich local semantic features and global context information.

For content veracity classification, in this paper, we model content veracity classification as a binary classification function:

$$f(n_i) \rightarrow y_i \tag{1}$$

using a set of labeled training news content data, i.e.,

$$D_{\text{train}} = \{(n_i, y_i)\}_{i=1}^{|D_{\text{train}}|} \tag{2}$$

y_i is the veracity label of the news article n_i, i.e., 0 for low-veracity content and 1 for high-veracity content. i is the i-th article, and $|D_{\text{train}}|$ is the total number of articles in the training dataset. We aim at learning the classification function:

$$f(n_i; \boldsymbol{\theta}) = \hat{y}_i \tag{3}$$

where $\hat{y}_i \in \{0, 1\}$ denotes the predicted probability of the article content being 1 (i.e., high-veracity content) and $\boldsymbol{\theta}$ is the learnable parameter vector. By minimizing the following objective function, our proposed model can predict the veracity of

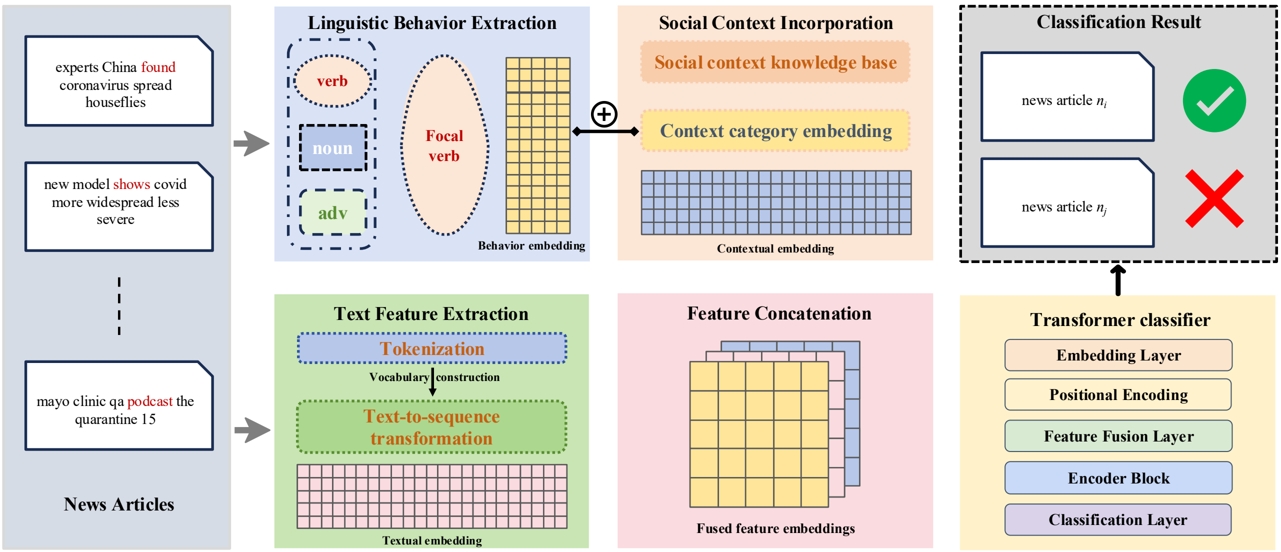

Figure 1: An illustration of our proposed L^{3}B pipeline

unseen content instances:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{|D_{\text{train}}|}\sum_{i=1}^{|D_{\text{train}}|}\Big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big] \tag{4}$$

3.2 Linguistic behavior extraction

To extract and represent the linguistic behaviors from article content, we define verbs as the indicators of linguistic behavior. For each article n_i, the verb set $\boldsymbol{V}_i$ is extracted using the spaCy toolkit: $\boldsymbol{V}_i = \{\boldsymbol{v}_i^1, \boldsymbol{v}_i^2, \ldots, \boldsymbol{v}_i^m\}$, where $\boldsymbol{v}_i^j$ is the j-th verb vector in n_i and m is the number of extracted verbs in article n_i. Technically, these extracted verbs are explored to explicitly represent linguistic behaviors inherent in n_i. To quantify the verb feature, we adopt TF-IDF to represent the importance of each verb.

$$\text{TF-IDF}(\boldsymbol{v}_i^j) = \text{TF}(\boldsymbol{v}_i^j, n_i)\times \log\left(\frac{|D_{\text{train}}|}{|\{n_k \in D_{\text{train}} : \boldsymbol{v}_i^j \in n_k\}| + 1}\right) \tag{5}$$

Where TF-IDF($\boldsymbol{v}_i^j$) is the frequency of verb $\boldsymbol{v}_i^j$ in news article n_i.

3.3 Social context incorporation

To incorporate social context information, we design a knowledge-based linking scheme to embed the social context features. Here, the Sentence-Transformer model (Reimers and Gurevych, 2019) is exploited to generate contextual embeddings and then access the similarity between extracted verbs $\boldsymbol{V}_i$ and contextual verbs in a predefined social context verb set $\boldsymbol{V}_{\text{context}}$. We adopt cosine similarity to quantify the semantic similarity between $\boldsymbol{V}_i$ and $\boldsymbol{V}_{context}$:

$$s_{ij} = \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_{\text{context}}^j}{\|\boldsymbol{v}_i\|\,\|\boldsymbol{v}_{\text{context}}^j\|} \tag{6}$$

where $\boldsymbol{v}_i \in \boldsymbol{V}_i$ is the embedding of extracted verbs in the content from article n_i. $\boldsymbol{v}_{\text{context}}^j \in \boldsymbol{V}_{\text{context}}$ is the embedding of j-th verb in the verb set $\boldsymbol{V}_{context}$. Then, the top-k embedding features from linked knowledge base with the highest s_{ij} values for each article n_i, i.e.,

$$\phi_{\text{k}}(n_i) = \text{Top}_k\left(\{s_{ij}\}_{j=1}^{k}\right) \tag{7}$$

Here, $k \in |\boldsymbol{V}_{\text{context}}|$, and $|\boldsymbol{V}_{\text{context}}|$ is the total number of verbs in the context verb set $\boldsymbol{V}_{\text{context}}$.

3.4 Feature fusion scheme

In this section, we introduce the feature fusion scheme in the proposed model L^{3}B. Multifaceted features are fused before content veracity classification, which includes:

- **Content feature** ϕ_{c}: extracted from raw content text.
- **Behavior feature** ϕ_{b}: TF-IDF vectors of $\boldsymbol{V}_i$ derived from the extracted verbs from article content.
- **Context feature** ϕ_{k}: embedding features from social context knowledge linking.

Combining these feature representations for each article n_i, we form the fused feature embeddings as:

$$\boldsymbol{\phi}_i = [\boldsymbol{\phi}_{\mathrm{c}}, \boldsymbol{\phi}_{\mathrm{b}}, \boldsymbol{\phi}_{\mathrm{k}}], \tag{8}$$

where $\boldsymbol{\phi}_i \in \mathbb{R}^{d_\phi}$ and d_ϕ is the dimension of the fused feature vector.

3.5 Transformer-based classifier

To classify content veracity, the fused features $\boldsymbol{\phi}_i$ are fed into a Transformer-based classifier, where a multi-head self-attention mechanism is employed to capture the local features and global relations among different features. More specifically, multi-head attention is derived as follows:

$$\mathrm{Multihead}(\boldsymbol{\phi}_i) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H)\boldsymbol{W}, \tag{9}$$

where each head is computed by:

$$\begin{aligned} \mathrm{head}_h &= \mathrm{Attention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h) \\ &= \mathrm{softmax}\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^T}{\sqrt{d_k}}\right)\boldsymbol{V}_h \end{aligned} \tag{10}$$

In which $\boldsymbol{Q}_h = \boldsymbol{Z}\boldsymbol{W}_h^Q$, $\boldsymbol{K}_h = \boldsymbol{Z}\boldsymbol{W}_h^K$, and $\boldsymbol{V}_h = \boldsymbol{Z}\boldsymbol{W}_h^V$ are the query, key, and value matrices, respectively, and $\boldsymbol{W}$, $\boldsymbol{W}_h^Q$, $\boldsymbol{W}_h^K$, and $\boldsymbol{W}_h^V$ are the learnable parameter matrices. h is the h-th attention head, and $\boldsymbol{H}$ is the total number of attention heads used in the multi-head attention mechanism.

Following the Transformer encoder, the final hidden representation is passed through a fully connected (FC) layer to generate the prediction logits:

$$\hat{y}_i = \sigma(\mathrm{FC}(\mathrm{Transformer}(\boldsymbol{\phi}_i))) \tag{11}$$

where σ denotes the sigmoid activation function.

3.6 Content veracity classification

Finally, $\hat{y}_i$ is obtained by setting a threshold of 0.5 to the predicted label probability, i.e.,

$$y_i^* = \begin{cases} 0 & \text{if } \hat{y}_i < 0.5, \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

Here, y_i^* is the predicted label of news content. We employ the cross-entropy loss to optimize the training process, using the Adam optimizer, mixed-precision training, and early stopping to effectively reach convergence.

4 Experiments

In this section, we conduct extensive comparison experiments on four public datasets collected from real-world scenarios, and experimental results demonstrate that our models have superior performance and efficiency than most tested models. We first introduce the experimental setup, including the datasets and tested models. Then, we report the experiment results and analyze these results for further exploration. Furthermore, the ablation study shows the modules contributing to the performance improvement.

4.1 Experimental setup

Datasets. For conducting the extensive experiments, we use four datasets to broadly test our model, compared to other advanced models, including health datasets (MM COVID (Li et al., 2020) and RoCOVery (Zhou et al., 2020)), news content dataset (LIAR (Wang, 2017)), and multi-domain dataset (MC Fake(Min et al., 2022)). The statistics of these four datasets are listed in Table 1.

Experimental Models. To fairly conduct the comparison experiments, we compared our proposed model with five other models, only using textual data without any additional modalities. The tested methods include a CNN-based model (Kim, 2014), a GCN-based model (Yao et al., 2018), HAN (Yang et al., 2016), BERT (Devlin et al., 2019), and HeteroSGT (Zhang et al., 2024). More specifically, the CNN-based model employs CNN layers to extract text features from article content and then uses the extracted features to classify content veracity. The GCN-based model explores the weighted graph built on news articles, which uses a GCN for content classification. HAN applies word-level and sentence-level features in news content for content veracity classification. BERT is

Dataset	# Label 0	# Label 1	# Total	Avg. Length (words)
MM COVID	1,888	1,162	3,048	25
RoCOVery	605	1,294	1,899	500
LIAR	2,507	2,053	4,560	17
MC Fake	2,671	12,621	15,292	300

Table 1: Statistics of the datasets used in our experiments.

Table 1: Detection performance on five datasets (best in red, second-best in blue).

Dataset	CNN		GCN		BERT		HAN		HeteroSGT		L^3B	
	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre	Acc	Pre
MM COVID	0.582±0.035	0.478±0.170	0.717±0.156	0.735±0.236	0.730±0.093	0.727±0.094	0.855±0.005	0.854±0.005	0.925±0.004	0.921±0.006	0.902±0.116	0.902±0.110
ReCOVery	0.658±0.011	0.460±0.104	0.718±0.037	0.691±0.178	0.682±0.030	0.441±0.213	0.722±0.021	0.462±0.197	0.909±0.002	0.902±0.002	0.879±0.017	0.865±0.028
MC Fake	0.825±0.001	0.544±0.156	0.724±0.138	0.516±0.169	0.827±0.006	0.713±0.271	0.825±0.005	0.463±0.098	0.883±0.002	0.812±0.003	0.887±0.051	0.827±0.016
LIAR	0.546±0.019	0.432±0.181	0.487±0.039	0.493±0.047	0.537±0.007	0.513±0.017	0.546±0.025	0.493±0.036	0.581±0.002	0.580±0.003	0.605±0.041	0.601±0.045
Dataset	Rec	F1	Rec	F1	Rec	F1	Rec	F1	Rec	F1	Rec	F1
MM COVID	0.547±0.039	0.474±0.101	0.685±0.178	0.621±0.184	0.722±0.101	0.720±0.103	0.854±0.006	0.853±0.005	0.915±0.005	0.918±0.005	0.898±0.142	0.900±0.132
ReCOVery	0.501±0.020	0.422±0.107	0.609±0.102	0.516±0.021	0.722±0.081	0.416±0.032	0.506±0.002	0.457±0.013	0.865±0.006	0.893±0.003	0.843±0.203	0.854±0.208
MC Fake	0.501±0.002	0.455±0.004	0.552±0.169	0.470±0.039	0.502±0.001	0.451±0.002	0.500±0.004	0.453±0.001	0.762±0.002	0.783±0.003	0.700±0.099	0.738±0.109
LIAR	0.502±0.005	0.377±0.049	0.494±0.029	0.423±0.055	0.510±0.012	0.483±0.014	0.502±0.018	0.445±0.053	0.575±0.002	0.571±0.003	0.595±0.037	0.595±0.037

Table 2: **Classification performance on four datasets** (best in red, second-best in blue).

a transformer-based language model, similar to our transformer-based classifier, where we explore BERT to classify false content (i.e., low-veracity content). HeteroSGT explores the heterogeneous subgraph transformer to classify articles via the heterogeneous graph.

4.2 Experiment Settings

Model Configuration. Our detection pipeline employs a transformer-based classifier, which effectively integrate the textual, linguistic behavior, and contextual features. Each input token is represented by a 128-dimensional embedding vector. In the transformer-based classifier module, our model consists of multiple stacked transformer encoder layers with a multi-head attention scheme. In the feature fusion function, we pool the transformer outputs and concatenate these features with content features, verb-based linguistic behavior features, and social context embeddings using semantic linking. For the fully-connected layer, we employ ReLU as an activation function and set dropout regularization to 0.1. The final output layer with a sigmoid function is designed to provide the probability scores indicating the likelihood of the content being labeled 1 (i.e., high veracity). Here, we use Adam optimizer with learning rate 1×10^-5 and cross-entropy loss to train our model, where we employ the mixed precision training and early-stopping to tune the hyperparameters. For training and testing our proposed model, we split all the datasets into train, validation, and test datasets using a ratio of 80%, 10%, and 10%, respectively. To validate the generalizability of tested methods, we perform 10 rounds of tests with random seeds for each model and then record the averaged results and standard deviation. Here, all the experiments are conducted on 1 NVIDIA A100 GPU with 64G RAM.

Evaluation Metrics. We quantitatively evaluate our model's performance compared to the other five tested models, using classification metrics such as accuracy (Acc), Macro-precision (Pre), Macro-F1 (F1), and Macro-recall (Rec).

4.3 Experimental Results

In Table 2, we report the experimental results of all the tested models across the four datasets. From Table 2, one can see that our model achieves superior performance across all the metrics on the LIAR dataset, and suboptimal performance on the datasets MM COVID, ReCOVery, and MC Fake. It shows that the linguistic behavior features can improve the model performance and have a significant impact on classifying content veracity. Additionally, we can see that our model achieves higher recall values on all four datasets, typically on the LIAR dataset. A higher recall indicates that less low-veracity content is missed when classifying the high- and low-veracity content. Furthermore, it should be noted that our model has robust and consistent performance across all the datasets, compared with other tested models.

For the five comparison models, CNN has poor performance on all the datasets, which may result from its fixed convolutional kernels. Due to these kernels focusing on local features, the global features or dependencies might not be effectively explored in news articles and social contexts. GCN presents different results across multiple datasets and receives better detection accuracy on MC Fake dataset. HAN and BERT are transformer-based models with attention mechanisms, and thus, the performance is comparable between HAN and BERT. Though HeteroSGT achieves optimal results on most datasets due to its subgraph structure, it still drops performance by 0.4% on Acc and 1.5% on Pre, 2.4% on Acc and 2.1% on Pre, respectively, compared to our proposed model on MC fake and LIAR datasets. Typically, on the MM COVID dataset, our model achieves consistent performance across seeds, with a low standard deviation (± 0.116).

Table 3: Ablation results on the ReCOVery dataset. We report Acc, Pre, Rec, and F1. ϕ_c: content (TF-IDF); ϕ_b: behavioral (verbs); ϕ_k: knowledge (context features).

Model Variant	Acc	Pre	Rec	F1
Full Model ($\phi_c + \phi_b + \phi_k$)	**0.857**	**0.861**	**0.829**	**0.840**
No Content ($\phi_b + \phi_k$)	0.813	0.835	0.763	0.779
No Knowledge ($\phi_c + \phi_b$)	0.808	0.803	0.776	0.785
Raw Text Only (ϕ_c)	0.783	0.767	0.770	0.769

4.4 Ablation Study

We conducted an ablation study on the ReCOVery dataset to evaluate the performance of three feature modules in our model: content feature (ϕ_c), behavioral feature (ϕ_b), and knowledge-based context feature (ϕ_k). From Table 3, we can see that the L^3B model incorporating all three modules achieved the best performance, i.e., accuracy of 0.857, Pre of 0.861, Rec of 0.829, and F1 score of 0.840. When removing the content features, it leads to the largest drop in recall (from 0.829 to 0.763) and a significant drop in F1 score (to 0.779), showing the importance of capturing nuanced textual features. Without knowledge features (ϕ_k), it also reduces the overall performance, such as F1-score from 0.840 to 0.785, indicating the significance of external social context in contextually grounding content for extracted verbal features from news articles. Additionally, using only textual content (ϕ_c), our model achieves an accuracy of 0.783, demonstrating the performance of the transformer-based baseline model. From these results, it can be seen that behavioral indicators and contextual features provide complementary gains beyond content-based representations alone.

5 Conclusion

Low-veracity content significantly disrupts content quality and integrity, and therefore, it's increasingly important to develop efficient and robust content classification models. In this work, we introduce the L^3B, a behavior-aware classification model, which leverages linguistic behaviors to classify high- or low-veracity content, alongside textual and contextual data. Compared to traditional ML-based approaches that rely heavily on content data, L^3B effectively mitigates the limitations of data dependency and bias through multi-faceted feature extraction. Firstly, the verb features are extracted from news articles as linguistic behavioral features. Then, a knowledge-based linking scheme is introduced to align the extracted verbal features with those derived from social context categories and further refine the behavioral features. Finally, the text, behavior, and context features are fused and fed into a transformer-based classifier to flag the content veracity. Experimental results show that L^3B outperforms most advanced classification models both in accuracy and generalizability, indicating the merits of integrating linguistic behavior features and behavior-context features into the classification and detection frameworks of content veracity on social media platforms.

Limitations

Though our proposed L^3B framework has superior performance in the content veracity classification task by integrating textual, behavioral, and contextual features, several limitations remain. First, our model demands the verbal features, which leads to poor performance in verb-sparse stances. Secondly, we incorporate the predefined knowledge base to model the social context for extracted verbal features, lacking adaptability to newly emerging social topics and content styles. Additionally, L^3B does not incorporate social credibility or propagation patterns into the content classification pipeline. Finally, our proposed model is restricted to text data and can not analyze multimodal content, such as images or videos.

For future studies, we aim to incorporate behavior credibility and reliability into the content veracity pipeline and also plan to explore more inherent features in news content, further improving model performance and generalizability across diverse datasets. In addition, we plan to introduce the multi-modality modules in the L^3B pipeline to capture text and image features and identify the consistency between textual and visual features for content veracity detection and classification of multimodal data.

References

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):1–29.

Sajjad Ahmed, Knut Hinkelmann, and Flavio Corradini. 2022. Combining machine learning with knowledge engineering to detect fake news in social networks-a survey. *ArXiv*, abs/2201.08032.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2021. Explainable tsetlin machine framework for fake news detection with credibility score assessment. *ArXiv*, abs/2105.09114.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 549–556.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7059–7069.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2025. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23760–23768.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2051–2055, New York, NY, USA. Association for Computing Machinery.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 81–89.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Dongqi Fu, Yikun Ban, Hanghang Tong, Ross Maciejewski, and Jingrui He. 2022. Disco: Comprehensive and explainable disinformation detection. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4848–4852.

Amira Ghenai and Yelena Mejova. 2018. Fake cures: user-centric modeling of health misinformation in social media. In *Proceedings of the ACM on Human-Computer Interaction*, volume 2, pages 1–20. ACM New York, NY, USA.

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Survery*, 53(4).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Syed Mustafa Haider Rizvi, Ramsha Imran, and Arif Mahmood. 2025. Text classification using graph convolutional networks: A comprehensive survey. *ACM Computing Survery*, 57(8).

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Ammar Ismael Kadhim. 2019. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52:273–292.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Junehyung Kim and Sungjae Hwang. 2024. All you need is attention: Lightweight attention-based data augmentation for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12866–12873, Miami, Florida, USA. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Guoyi Li, Die Hu, Zongzhen Liu, Xiaodan Zhang, and Honglei Lyu. 2025. Semantic reshuffling with LLM and heterogeneous graph auto-encoder for enhanced rumor detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8557–8572, Abu Dhabi, UAE. Association for Computational Linguistics.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *ArXiv*, abs/2011.04088.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4821–4830, Hong Kong, China. Association for Computational Linguistics.

Chao Liu, Xinghua Wu, Min Yu, Gang Li, Jianguo Jiang, Weiqing Huang, and Xiang Lu. 2019. A two-stage model based on bert for short fake news detection. In *Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part II 12*, pages 172–183. Springer.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Qianli Ma, Zhenxi Lin, Jiangyue Yan, Zipeng Chen, and Liuhong Yu. 2020. MODE-LSTM: A parameter-efficient recurrent network with multi-scale for sentence classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6705–6715, Online. Association for Computational Linguistics.

Valeria Mazzeo, Andrea Rapisarda, and Giovanni Giuffrida. 2021. Detection of fake news on covid-19 on web search engines. *Frontiers in Physics*, 9:685730.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1148–1158, New York, NY, USA. Association for Computing Machinery.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Computing Survery*, 54(3).

Rahul Mishra and Vinay Setty. 2019. Sadhan: Hierarchical attention networks to learn latent aspect embeddings for fake news detection. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 197–204, New York, NY, USA. Association for Computing Machinery.

Kashyap Popat. 2017. Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 735–739, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. 2019. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6940–6948.

Isabel Segura-Bedmar and Santiago Alonso-Bartolome. 2022. Multimodal fake news detection. *Information*, 13(6):284.

Lanyu Shang, Yang Zhang, Bozhang Chen, Ruohan Zong, Zhenrui Yue, Huimin Zeng, Na Wei, and Dong Wang. 2024. Mmadapt: A knowledge-guided multi-source multi-class domain adaptive framework for early health misinformation detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4653–4663, New York, NY, USA. Association for Computing Machinery.

Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 101–102, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Chongyang Shi, Yijun Yin, Qi Zhang, Liang Xiao, Usman Naseem, Shoujin Wang, and Liang Hu. 2023. Multiview clickbait detection via jointly modeling subjective and objective preference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11807–11816, Singapore. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: a data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12nd ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2020. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 436–439, New York, NY, USA. Association for Computing Machinery.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1):1–13.

Xing Su, Jian Yang, Jia Wu, and Yuchen Zhang. 2023. Mining user-aware multi-relations for fake news detection in large scale online social networks. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, WSDM '23, page 51–59, New York, NY, USA. Association for Computing Machinery.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, Online. Association for Computational Linguistics.

Xian Teng, Yu-Ru Lin, Wen-Ting Chung, Ang Li, and Adriana Kovashka. 2022. Characterizing user susceptibility to covid-19 misinformation on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1005–1016.

Kjerstin Thorson, Emily Vraga, and Brian Ekdale. 2010. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society*, 13(3):289–313.

Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A Butt. 2021. What social media told us in the time of covid-19: a scoping review. *The Lancet Digital Health*, 3(3):e175–e194.

Jens Van Nooten and Walter Daelemans. 2025. Jump to hyperspace: Comparing Euclidean and hyperbolic loss functions for hierarchical multi-label text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4260–4273, Abu Dhabi, UAE. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 275–284, New York, NY, USA. Association for Computing Machinery.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3367–3378, New York, NY, USA. Association for Computing Machinery.

You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. In *Proceedings of the VLDB Endowment*, volume 7, page 589–600.

Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. Fusing label embedding into BERT: An efficient improvement for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750, Online. Association for Computational Linguistics.

Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5705–5717, Singapore. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *ArXiv*, abs/1809.05679.

Jungmin Yun, Mihyeon Kim, and Youngbin Kim. 2023. Focus on the core: Efficient attention via pruned token compression for document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13617–13628, Singapore. Association for Computational Linguistics.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. 2024. Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10467–10484, Miami, Florida, USA. Association for Computational Linguistics.

Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 603–612, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yuchen Zhang, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2024. Heterogeneous subgraph transformer for fake news detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1272–1282, New York, NY, USA. Association for Computing Machinery.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3205–3212, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ArXiv*, abs/1906.04210.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Juergen Kurths. 2024. Propagation structure-aware graph transformer for robust and interpretable fake news detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4652–4663, New York, NY, USA. Association for Computing Machinery.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Survery*, 51(2).