Unveiling Deception: A Novel Framework Integrating BERT and Vocabulary Graph Convolution Networks for Fake News Detection

Anonymous ACL submission

Abstract

False information, commonly known as fake news, pertains to deceptive and incorrect content disseminated in news format on the internet. On one hand, it possesses the potential to influence the outcome of exit polls by tarnishing the reputation of a deserving candidate. On the other hand, it can contribute to the instigation of communal violence. In this paper, we underscore the distinct linguistic patterns existing between fake and real news. We introduce a novel framework for identifying fake news, involving the fusion of two separate vocabulary graphs derived from the lexical spaces of both fake and real news. Essentially, our approach integrates the strengths of BERT with these two distinct vocabulary graphs. Further, Vocabulary Graph Convolution Networks generate document representation for classification. We report the F1 score of 94.94 and 98.44 on the Recovery and Multi-Source Fake datasets, respectively. Results show that our approach outperformed eight other deep neural frameworks.

1 Introduction

Fake news denotes the dissemination of inaccurate or deceptive information, frequently having concealed motives that can disrupt social harmony, manipulate public opinion, or demean specific genders or communities. The proliferation of social media and the ease of information dissemination have significantly exacerbated the prevalence of fake news. In light of this pressing issue, there is an imperative need to establish a robust and high-performance fake news detection system. This paper introduces a novel approach aimed at effectively distinguishing between fake and real news articles.

Owing to its significant impact on society, the detection of fake news has emerged as a crucial research area within natural language processing. Advancements from manually crafted features to uncovering hidden features through deep neural networks have contributed to effective solutions for fake news detection. Recent research has leveraged pretrained language models (PLMs), such as BERT, to comprehend news content and generate representations (Zhang et al., 2021; Kaliyar et al., 2021; Mosallanezhad et al., 2022; Hu et al., 2023). Notably, fine-tuning BERT has proven to yield superior performance in detecting fake news when compared to larger language models like GPT 3.5 (Hu et al., 2024). Despite the improvements offered by PLMs, their limitations in domain knowledge and task-specific capabilities pose challenges for fake news detectors.

BERT determines document representations based on the significance or attention given to words in the document. In the context of fake news detection, it is paramount to consider the relationships between words rather than solely focusing on individual words themselves. In our research, we aim to overcome the limitations of the BERT Model by integrating it with both Real News Vocabulary Graph and Fake News Vocabulary Graph. Creating distinct and independent representations of text features for fake and real news allows us to capture the intrinsic characteristics of fake and real news word spaces individually. In contrast, graph neural networks have the capability to comprehend global contextual relations among words in a corpus. Therefore, our proposed approach involves learning two separate vocabulary graph representations specific to fake and real word spaces for a given input instance. Further, Vocabulary Graph Convolution Networks generate the final document representation for classification. This paper makes the following notable contributions.

1. We present a novel technique for detecting fake news by integrating vocabulary graphs learned from both fake and genuine word spaces. Essentially, it captures insights from the distinct linguistic patterns found in both fake and real contexts.

2. Our model produce a weighted F1 score of 94.94 and 98.44 on Recovery and Multi Source Fake datasets. Notably, our model outperformed eight other deep learning configurations, including state-of-the-art (Devlin et al., 2019), across the datasets. This underscores the robustness and efficacy of our approach in comparison to existing techniques.

The rest of the paper is organized as follows. Section 2 presents the literature review. Section 3 elaborates the methodology. Section 4 describes the experimental setup. Section 5 demonstrates the results. Section 6 concludes the paper.

2 Literature

Initially, fake news detection strategies relied on analyzing linguistic features in news, such as writing styles and consistency (Potthast et al., 2018). In more recent developments, the emphasis has shifted towards creating models capable of understanding relational structures between sentences within news articles, as genuine news often exhibit strong connections between sentences, a characteristic often absent in fake news (Karimi and Tang, 2019). Vaibhav et al. (Vaibhav et al., 2019) introduced a model that constructs relational graphs from sentences, where sentences serve as nodes, and the weights on edges represent the similarity between sentences. The model learns document representations based on sentence interactions. Kim and Ko (Kim and Ko, 2021) developed a graph based on the relationships between sentences, utilizing summary information as a crucial aspect of document content. Ghanem et al. (Ghanem et al., 2021) proposed a model for detecting fake news based on the flow of information within news text.

More recently, researchers have explored the extraction of relationships between words through spectral graphs (Kipf and Welling, 2016). Yao et al. (2018) were first in applying Graph Convolutional Networks to text classification tasks, demonstrating state-of-the-art performance. They constructed a graph from text based on word co-occurrence and document-word relationships. Following a similar trajectory, Zhang and Zhang (2020) and Huang et al. (2019) proposed models to generate graphs from text that are less memory-intensive than Yao et al.'s (2018) approach.

Lin et al. (2021) integrated BERT with TextGCN by initializing document embeddings with BERT Document Embeddings. Lu et al. (2020) combined Vocabulary GCN word embeddings with BERT embeddings using a transformer self-attention mechanism. Our work aligns with this line of research, specifically utilizing spectral relations among words in the corpus for fake news detection. Our approach involves generating and integrating two independent vocabulary graphs learned from both fake and genuine word spaces, capturing linguistic patterns found in both fake and real contexts independently.

3 Methodology

In this paper, we have proposed a novel setup that combines Vocabulary GCN (VGCN) with BERT for fake news detection.

3.1 Vocabulary Graph

In the vocabulary graph, the word is the node and the edge is decided using Normalized Point-wise Mutual Information (NPMI). To capture NPMI information, we slide a fixed size window on the text. The NPMI value of word i and word j is calculated as follows:

$$NPMI(i,j) = -\frac{1}{logp(i,j)} * \frac{p(i,j)}{p(i)p(j)} \quad (1)$$

where, $p(i,j) = \frac{\#w(i,j)}{\#w}, p(i) = \frac{\#w(i)}{\#w}$, $\#w(i,j)$ represents the number of windows containing the word i, and the word j. Here, $\#w$ is the total number of sliding windows. The range of NPMI values is -1 to 1. We consider the edge between words with a positive NPMI value (Lu et al., 2020).

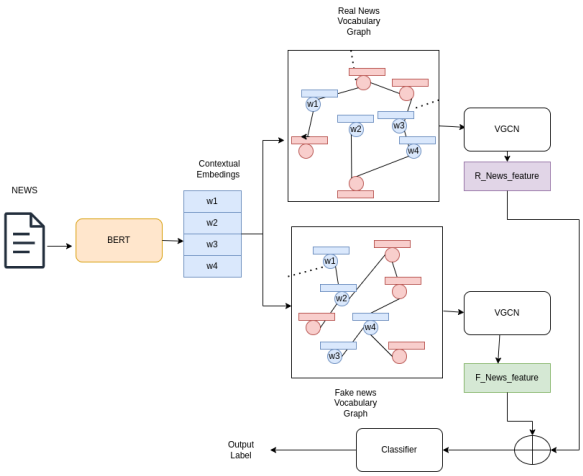

Figure 1: **Architecture of the Proposed Model**

3.1.1 Real News Vocabulary Graph

The Real News corpus is created using Real news articles that are included in the training dataset. The construction of the real news vocabulary graph is based on the NPMI using the real news corpus.

3.1.2 Fake News Vocabulary Graph

The fake news corpus is created using fake news articles that are included in the training data set. The creation of the Vocabulary Graph for fake news is built using NPMI considering the fake news corpus.

3.2 Vocabulary Graph Convolutional Networks (VGCN)

The single convolution layer of GCN for the new representation is calculated as follows.

$$H = \widetilde{A}XW \tag{2}$$

where, $X \in R^{n \times m}$ is the input matrix with n nodes and m dimensions of the features, $W \in R^{n \times h}$, $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, where, $D_{ii} = \sum_j A_{ij}$. A is the adjacency matrix of graph $G = (N, E)$, where N is in a set of nodes and E is a set of edges (Lu et al., 2020). Further, we use the following equation to compute document embeddings in VGCN:

$$G_{meg} = \sigma(X_{mev}\tilde{A}_{vv}W_{vh})W_{hg} \tag{3}$$

Every row of X_{mev} is the extended word embeddings of BERT (padding zero vectors for the non-appeared words in a sentence to the vocabulary size). Here, m is the batch size, word embedding size is e, and v is the vocabulary size. G_{meg} is the VGCN graph embedding of size g.

3.3 Proposed Model

Given a text document, it is first tokenized into a sequence of sub-word tokens and fed into the BERT model to obtain its contextual embeddings X_{met}, here $|T|$ represents number of sub-word tokens. We extend the X_{met} to X_{mev} by padding zero vectors for the non-appearing words in a sentence to the vocabulary size. Figure 1 presents the architecture of the model. The $RA_{v \times v}$ is the normalized adjacency matrix of real news vocabulary graph, m is the batch size, v is the vocabulary size, and g is graph embeddings. RG_{meg} (Equation 4) is graph (g) embeddings of size e with batch size m.

$$RG_{meg} = \sigma(X_{mev}CA_{vv}W_{vh})W_{hg} + b_g \tag{4}$$

$FA_{v \times v}$ is the normalized adjacency matrix of the fake news vocabulary graph, where v represents the size of the vocabulary. FG_{meg} (Equation 5) is graph (g) embeddings of size e with batch size m.

$$FG_{meg} = \sigma(X_{mev}SA_{vv}W_{vh})W_{hg} + b_g \tag{5}$$

We perform concatenation of the RG_{meg} and FG_{meg} to obtain the final document representation (D_{2meg}, Equation 6). $\oplus$ is the vector concatenation operation.

$$D_{2meg} = RG_{meg} \oplus FG_{meg} \tag{6}$$

Finally, the classification is performed by the output layer having $sigmoid$ activation function. This maps document representation D_{2meg} to a class label. Where, label $\in$\{Real or Fake\}.

4 Experimental Setup

We employed BERT embeddings to initialize the nodes of the vocabulary graphs, BERT-RF-VGCN (see Table 1). The hidden layer size is 156 and $g = 1$, with hyper-parameters set as follows: epoch $= 10$, dropout $= 0.2$, batch size $= 8$. Optimization is performed using BERTAdam with a learning rate of $1e - 5$. We assess performance using the weighted average F1-score (WF1), utilizing the cross-entropy loss function for all reported models.

4.1 Dataset

Results for all models are reported on the following datasets. **Multi Source Fake (Ghanem et al., 2021)**: This dataset comprises 5994 real and 5403 fake news articles. The test portion includes 689 fake news articles and 1000 real news articles. **Recovery Dataset (Zhou et al., 2020)**: This dataset contains 2049 news articles related to COVID-19, where 1364 are real and 665 are fake news articles. We used 80% of the dataset for training, 10% for validation and another 10% for testing. To account for the class imbalance, we present the results using the weighted F1 score.

4.2 Baseline Models

We extensively validated our hypothesis by comparing our approach (Section 3.3) with eight baseline models, *viz.*, Bi-LSTM (Graves and Schmidhuber, 2005), EANN-Text (Wang et al., 2018), TextGCN (Yao et al., 2018), TensorGCN (Liu et al., 2020), BERT (Devlin et al., 2019), VGCN-BERT (Lu et al., 2020), Fake-Flow (Ghanem et al., 2021), BERT with Pref-FEND (Sheng et al., 2021).

S. No.	Model	Recovery	Multi Source Fake
1	Bi-LSTM (Graves and Schmidhuber, 2005)	81.79	94.44
2	EANN-Text (Wang et al., 2018)	80.44	94.83
3	TextGCN (Yao et al., 2018)	70.94	89.71
4	TensorGCN (Liu et al., 2020)	87.24	91.93
5	BERT (Devlin et al., 2019)	92.50	96.80
6	VGCN-BERT (Lu et al., 2020)	91.54	96.67
7	Fake-Flow (Ghanem et al., 2021)	–	96
8	BERT with Pref-FEND (Sheng et al., 2021)	90.23	97.15
9	BERT with RF-VGCN (Our Approach)	**94.94**	**98.44**

Table 1: Weighted F1 (WF1) Score on *Recovery* and *Multi Source Fake* datasets

5 Results

We extensively evaluated the effectiveness of our approach, comparing our system with eight other fake news detection systems across two datasets. The weighted F1 (WF1) scores for the fake news detection task are outlined in Table 1. Rows 1,3 and 4 in Table 1 demonstrate that the LSTM model, when initialized with contextual embeddings, outperforms the graph-based model initialized with Glove word embeddings. This emphasizes the significance of contextual embeddings (Devlin et al., 2019) in the fake news detection task. The VGCN-BERT model (Lu et al., 2020) has demonstrated state-of-the-art performance in text classification tasks; however, in the context of fake news detection, it downgrades BERT's performance. The decline in performance can be attributed to the utilization of a single vocabulary, rendering the model unable to distinguish between the linguistic patterns present in fake news and real news.

Conversely, the amalgamation of BERT with the Pref-FEND model (Sheng et al., 2021) improves BERT's performance on the *Multi Source Fake* dataset, which spans multiple domains. While, there is a decline in performance within the health (*Recovery*) domain. Our observations suggest that the *Recovery* dataset lacks an adequate number of instances to gather information for a heterogeneous dynamic graph as per Pref-FEND model's settings.

BERT Vs Our Approach While BERT demonstrated performance scores of 92.50 and 96.80 on the Recovery and Multi Source FAke datasets, our approach, combining two distinct vocabulary graphs with BERT, yielded weighted F1 scores of 94.94 and 98.44 for the same datasets.

Model	Recovery WF1	Multi Source Fake WF1
M1	**94.94**	**98.44**
M2	**92.80**	**98.04**
M3	**93.14**	**98.11**
M4	**92.50**	**96.80**

Table 2: Ablation Study

5.1 Ablation Study

In Table 2, **M1** presents our approach, while **M2** represents the BERT model with the Fake News Vocabulary Graph but without the Real News Vocabulary Graph. **M3** denotes the BERT model with the Real News Vocabulary Graph but without the Fake News Vocabulary Graph, and **M4** represents the pure BERT model. The study reveals that both the Real News Vocabulary Graph and Fake News Vocabulary Graph components contribute to enhancing BERT's performance individually. However, their combined representation surpasses individual improvements in performance.

6 Conclusion

This paper addresses a widespread social concern: the detection of fake news. Our hypothesis suggests that combining distinct vocabulary graph neural networks for fake and real word spaces introduces complementary information essential for accurately classifying between fake and real news. Moreover, we incorporate the capabilities of BERT for text classification alongside vocabulary graph neural networks. The results demonstrate that our approach surpasses the performance of eight other deep neural configurations designed for fake news detection.

Limitations

Limitation of our model is that it requires more GPU memory during training than BERT model. The memory size depends on the size of the vocabulary.

Ethics Statement

The work reported in the paper complies with the ACL Ethics Policy. The output of our work does not have any negative impact on society; it is for the good of society.

References

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. FakeFlow: Fake news detection by modeling the flow of affective information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 679–689, Online. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024)*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 116–125, Toronto, Canada. Association for Computational Linguistics.

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for 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 (EMNLP-IJCNLP)*, pages 3444–3450, Hong Kong, China. Association for Computational Linguistics.

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.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. 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 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Gihwan Kim and Youngjoong Ko. 2021. Graph-based fake news detection using a summarization technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3276–3280, Online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. BertGCN: Transductive text classification by combining GNN and BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462, Online. Association for Computational Linguistics.

Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. 2020. Tensor graph convolutional networks for text classification. *CoRR*, abs/2001.05313.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: Augmenting bert with graph embedding for text classification. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 369–382. Springer.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3632–3640, New York, NY, USA. Association for Computing Machinery.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern- and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM International Conference on*

Information & Knowledge Management, CIKM '21, page 1640–1650, New York, NY, USA. Association for Computing Machinery.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139, Hong Kong. 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.

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

Haopeng Zhang and Jiawei Zhang. 2020. Text graph transformer for document classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8322–8327, Online. Association for Computational Linguistics.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, page 3465–3476, 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*, pages 3205–3212.