Debiased Social Media Fake News Detection based on User Portraits Simulation

Anonymous ACL submission

Abstract

Social media fake news detection aims to detect fake news from platforms through online interaction data, which mainly consists of user posts and related comments. Through statistics, we found that the number of replies to posts depends largely on the time of posting, which we named temporal bias of data. Traditional methods focus on graph modeling to explore the potential structures among social texts, but ignore data bias. Although related methods based on large language models (LLMs) generate interactive comments and perform input enhancement, the generated information is uncontrollable and does not address data bias. In response, we propose a approach that uses LLMs to debias through data augmentation, named DUPS. The method first uses the LLM to analyze the user portraits, and then simulates the corresponding portrait to generate interactive comments, thereby reconstructing unbiased data. Experimental results on three datasets show that DUPS outperforms the current State-Of-The-Art approaches.

1 Introduction

With the vigorous development of social platforms, people are more inclined to express opinions or consult information on the Internet. However, the increasing information is accompanied by more fake news, which has caused negative impacts on our lives. Therefore, it is crucial to detect fake news automatically and precisely.

Considering the deceitful content of news, early research devoted to exploring text content to predict news authenticity. These studies focus on modeling additional features such as emotional signals (Giachanou et al., 2019; Zhang et al., 2021), writing style (Yang et al., 2021; Zhu et al., 2022), and text grammar or patterns (Azevedo et al., 2021; Horne and Adali, 2017) to improve the accuracy of fake news label prediction. However, for fake news that carefully tailored by compilers, it is difficult for the model to predict their labels based on the text content alone.

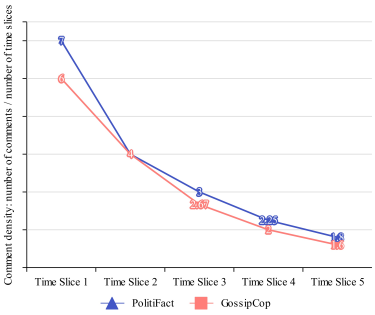


Figure 1: Both the average comments density in PolitiFact and GossipCop test set shows a downward trend. The horizontal axis is the time slice, which defaults to 12 hours. The vertical axis is the comment density, which is calculated as: total number of existing comments/number of time slices.

On the other hand, some work focused on using social networks to collect evidences as an important basis for predicting the authenticity of news. Some studies crawl information from authoritative lines as evidence and establish an benchmark (Augenstein et al., 2019). Due to the scarcity of authoritative information, more studies tend to use social wisdom: crawling text like user comments as evidences (Yuan et al., 2019; Yang et al., 2023). Although these works have achieved improvements by modeling user comments, we have observed that the density of comments under news in some benchmarks shows a rapid decline over time, as shown in Figure 1. On social media, comments are usually responses to the posts under news. The decreasing trend in comment density over time means that subsequent posts will receive less attention, which may cause useful but untimely posts to be ignored by the model. As an example from twitter shown in Figure 2, although the posts in the blue box can be used as a basis for fake news detection, they did not receive much attention from the social plat-

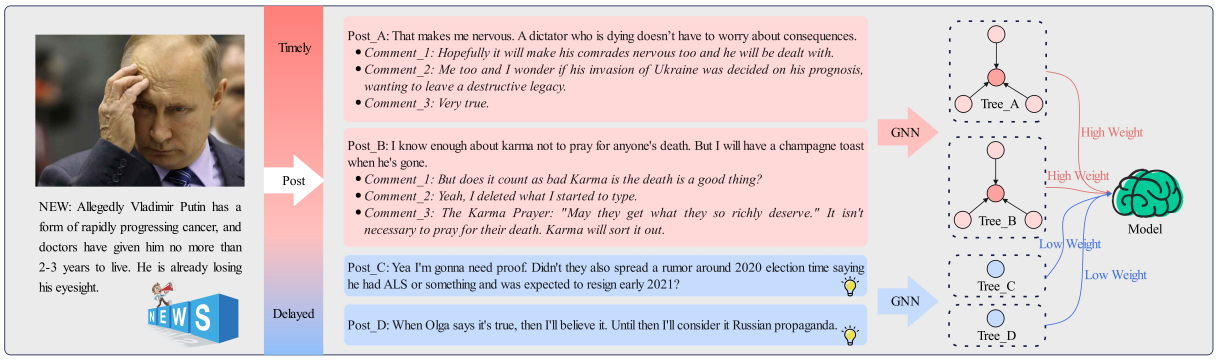


Figure 2: An example from twitter: A timely post may receive more comments, thus the graph centered on the post will be larger, making it more likely to obtain a higher weight when modeling.

form (no response was received) due to their late release. The big gap in the number of comments under posts at different times, which will be referred to as the temporal bias of posts below, is a common phenomenon in fake news detection on social media.

Recently, with the development and exploration of LLMs, its capabilities in role-playing (Shao et al., 2023) and personalized language generation (Woźniak et al., 2024) have been gradually explored. Considering that LLMs can simulate social user portraits to generate personalized comments, we propose **D**ebiased Social Media Fake News Detection based on **U**ser **P**rotraits **S**imulation. DUPS obtains debiased data through data augmentation to predict fake news labels, which can be divided into three steps: First, for existing comments, LLM is used to score and cluster based on five personal attributes to analyze user portraits; second, for posts with temporal bias, LLM is utilized to simulate user portraits and generate corresponding comments for debiasing; last, the debiased data is modeled to predict news labels.

Our contributions are summarized as follows:

1) Based on Big 5 Personality Traits (Lim, 2023), five personal attributes are set to characterize user portraits, which is more controllable and reasonable.

2) We proposed DUPS, which uses the LLM to analyze user portraits, then generates corresponding comments through simulation to remove the temporal bias of posts.

3) Experimental results on three datasets show that DUPS outperforms the current State-Of-The-Art approaches.

2 Related Work

2.1 Social Media Fake News Detection

Different from content-based methods, social media fake news detection approaches are dedicated to exploring the potential structures graph modeling, such as word relations, news dissemination process and social structure. Yao et al. (2018) proposed a method to construct a weighted graph using the words contained in news content and then apply a graph convolutional network to classify fake news. Similarly, Linmei et al. (2019) proposed a methodology, in which a heterogeneous graph attention network is used to construct a multi-relational graph for classification. Besides, Ma et al. (2018) and Bian et al. (2020) focus on capturing the features in terms of the fake news propagation by utilizing RNN and bi-directional GCN, respectively. Other works like Dou et al. (2021) and Su et al. (2023), which model the relations between news and users for fake news detection.

Although these works have achieved improvements, they are based on biased social data, which lowers the upper bound of methods.

2.2 LLM in Fake News Detection

As LLMs are deeply developed and mined, their reasoning and generation capabilities are used in various downstream tasks, including fake news detection. Refering to some work that using LLMs for generating chain of thought, Hu et al. (2024) took LLM as an Advisor in fake news detection task and distilled its knowledge into the small model. Based on the efficient retrieval and information integration capabilities of online LLMs, Li et al. (2024) took the LLMs as agents. From evidence-searching to decision-making, each agent collaborated to complete the task. In addition, some works focus on

detecting fake news that generated by LLMs. Chen and Shu (2024) conducted a systematic research on fake news generated by LLMs, including detection difficulty assessment and disinformation classification. Lucas et al. (2023) explored whether a large language model can detect fake news generated by other LLMs and found that it is feasible through special instructions. Although these works have explored or utilized large models in fake news detection, their integration with LLMs remains superficial.

Besides, Wan et al. (2024) proposed a method, which simulates the social user network through LLMs, by setting user portraits and network structures in advance. However, the user interact-network simulated in this work is not authentic and comprehensive, in which all user portraits and relationship are set in advance. Furthermore, it generates interaction networks only for the purpose of input enrichment, but does not target explicit issues such as data bias. In response, we propose an approach to simulate social networks based on real comments via LLMs, aiming for data debiasing.

3 Methodology

In this section, we first provide the problem statement of the task, followed by details of our method. The overall framework of DUPS is depicted in Figure 3. The figure shows that the implementation of this method can be divided into three steps, each of which is elaborated in detail in the following sections.

3.1 Problem Statement

Given a online news with $|X|$ word as $x = \{x_1, x_2, ..., x_{|X|}\}$, its posts list as $P = \{p_1, p_2, ..., p_{|P|}\}$ and relevant comments set as C. For post a, its comment set is denoted as $C_a = \{c_{a1}, c_{a2}, ..., c_{a|C_a|}\}$, where $C_a \in C$. Each news piece has a ground-truth label $y \in \{0, 1\}$, where 0 and 1 denote the new piece is fake and real, respectively. Social fake news detection aims to detect the anthenticity the online news through its posts and relevant comments set.

3.2 Obtaining User Portraits

As shown at the top of Figure 3, obtaining user portraits is a two-step pipeline process: using LLM to analyze personal attribute, and then getting portraits through clustering.

3.2.1 Personal Attribute Analysis

The Big 5 Personality Traits is a five-factor classification method for studying personality and has been widely used in various applications. Based on this approach, we use LLM to analyze the user personality of a given comment from five personal attributes and score each dimension (0-5 points, the higher the score, the stronger the attribute). Each personal attribute and corresponding explanation are as follows:

Openness: Refers to the degree of openness of thought. For example, "There are many possibilities for the future" and "Don't think about it" represent the open-minded and conservative factions respectively.

Optimism: Refers to the degree of optimism in attitude. For example, "I am looking forward to it" and "I am extremely disappointed" represent an optimistic and pessimistic attitude respectively.

Rationality: Refers to the stability of emotions. For example, "Not just because I like it" and "Wow, I really like it" represent rational and emotional expressions respectively.

Logic: Refers to the rigor of thinking. For example, "Because of the above arguments, it is very reasonable" and "It is unreasonable, no reason" represent logical and chaotic thinking respectively.

Affinity: Refers to the intensity of expression. For example, "This is really nonsense" and "This is not very reasonable" represent intense and euphemistic expressions respectively

3.2.2 User Portraits Clustering

Given a comment, the above personal attributes are scored using the LLM to obtain a five-dimensional vector. Referring to the method proposed by Zhang et al. (2023), we use LLM to cluster all scored comments into K categories (the value of K is set through experiments), and then summarize the crowd portraits of each category by the LLM as the user portrait of each comment under the corresponding cluster. The reason why the clustered crowd portrait is used as the user portrait of each comment in the group is that the number of personalities in real world are limited. Now each comment has a corresponding user portrait, denoted as U_C ($|U_C| = |C|$).

3.3 Comments Generation

As shown in the middle of Figure 3, comments generation is also a two-step pipeline process: determining whether to generate comments and it-

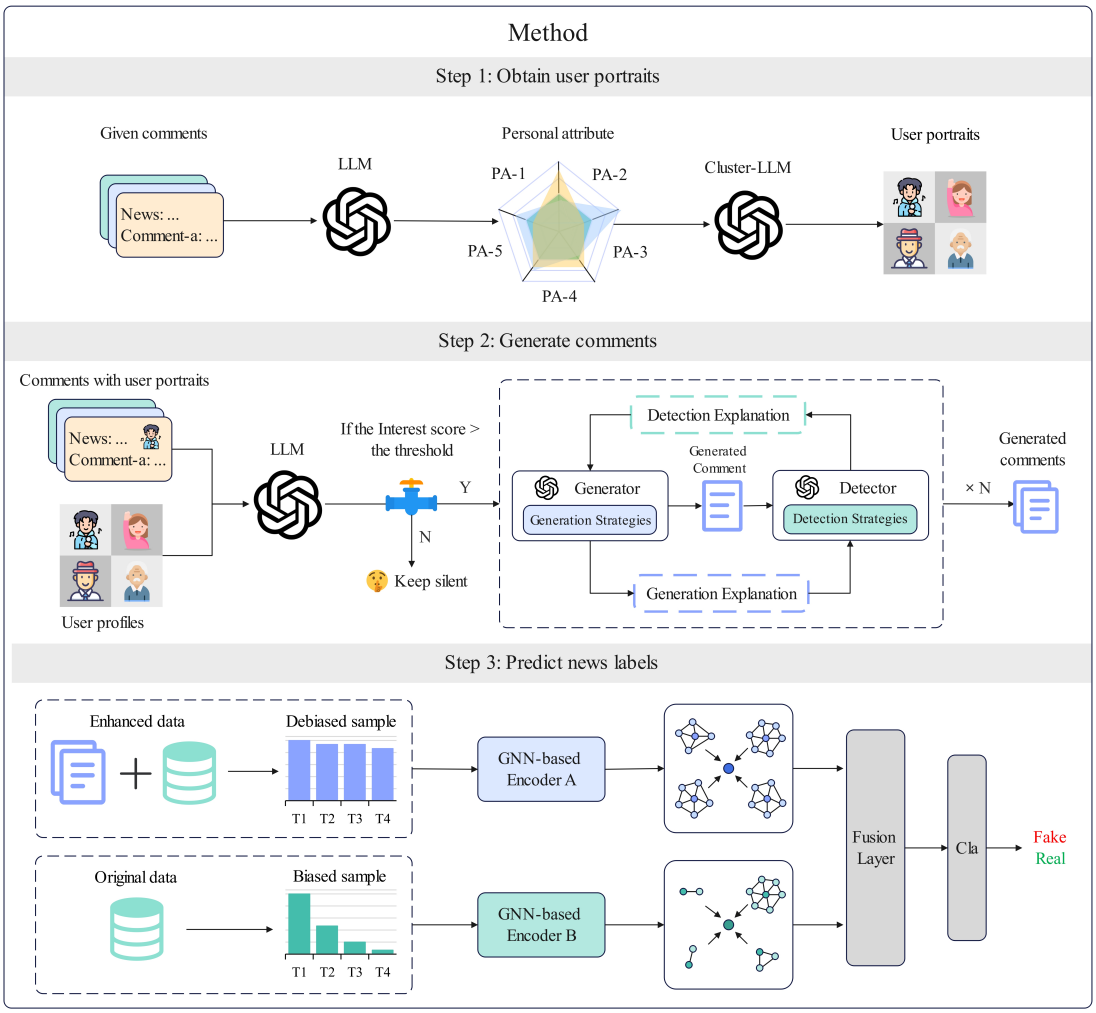


Figure 3: The overall framework of DUPS.

eratively optimizing the comments generated by simulation.

3.3.1 Determine whether to Generate Comments

Given a news X , its post list P and relevant comments set C, we perform the following steps:

1. Determine whether there is a time deviation in the posts in the list. If not, skip to next news, otherwise continue.

2. Search for posts with few comments, and save them in the list T.

3. For each t in the list T, use LLM to analyze whether each user portrait in the C set is interested in s and give a score (0-5 points, the higher the score, the more interested the user is).

4. The interest value score need to be compared with the threshold (set to 3 in the experiment). If it is higher than the threshold, the generation line continues, otherwise it remains silent.

When the interest score is high enough, it means that the user with a certain portrait is interested in the current post (which has few replies and needs to be debiased), and then it enters the generation line, which is elaborated in detail in the next section.

3.3.2 Iteratively Optimizing the Comments Generated

In order to make the generated comment closer to the simulated user portrait, we utilize LLM-GAN, which proposed by Wang et al. (2024), to iteratively optimize the comments. As shown in the right half of step 2 in Figure 3, we introduce a LLM generator M_G and a LLM detector M_D to perform adversarial training together.

Generator M_G: The goal of the generator M_G is to simulate a specific user portrait under a post to generate comment (c_G) and give the corresponding reasons (r_G) for generation. The generation process is as follows:

$$c_G, r_G = M_G(X, p, u_c, S_G) \quad (1)$$

Where X, p and u_c represent the given news and post, as well as the comments posted by the user portrait to be simulated, respectively. S_G is the generation strategy of the generator, which is initialized by Equation 2 and updated by Equation 3.

$$S_{G-init} = M_G(X, p, u_c) \tag{2}$$

$$S_G = M_G(X, p, u_c, r_D) \tag{3}$$

where r_D is the reasons for detection given by the detector M_D, which is elaborated below.

Detector M_D: The goal of the detector M_D is to detect whether a comment is posted by a user of a certain portrait based on the content of the comment (y_D) and give the corresponding reasons (r_D) for detection. The detection process is as follows:

$$y_D, r_D = M_D(X, p, u_c, S_D) \tag{4}$$

Where S_D is the detection strategy of the detector, which is initialized by Equation 5 and updated by Equation 6.

$$S_{D-init} = M_D(X, p, u_c, c_u) \tag{5}$$

$$S_D = M_D(X, p, u_c, r_G) \tag{6}$$

Where c_u denotes the real comment posted by the user with a certain portrait currently being detected.

Through adversarial training, the generator continuously optimizes its strategy to generate comments that are more consistent with the current simulated user portrait.

3.4 News Labels Prediction

Since the generated data, although close to user portraits, may not be effectively used to predict the task label, we combine real biased data to alleviate this problem. As shown at the bottom of Figure 3, news labels prediction consists of two pathways, taking the debiased and the original data as input, respectively. Then, the two parts of data are modeled through graph neural networks to obtain the corresponding representations, as shown in the Equation 7 and Equation 8.

$$E_{De} = GCN(T(X, P, C_{De}, \theta_1), \theta_2) \tag{7}$$

$$E = GCN(T(X, P, C, \theta_3), \theta_4) \tag{8}$$

where $T(*)$ represents the transformer encoder and C_{De} denotes the debiased comments set. $\theta_1 - \theta_4$ are represent the parameter sets of the corresponding models.

The graph representations of debiased and original data are combined through a fusion layer to obtain a feature vector, which serves as the input of the classification layer to output the predicted label:

$$V_{fusion} = \left[E_{De}; E; E_{De} - E; E_{De} \cdot E\right] \tag{9}$$

$$P(y|x) = Softmax(Linear(V_{fusion})) \tag{10}$$

where $[;]$ denotes concatenation of vectors.

4 Experiments

In this section, we first introduce the experimental settings, including: datasets, experimental details and baselines, then present the results.

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on PoliticFact (Shu et al., 2019b), GossipCop (Shu et al., 2019b) and CoAID (Li et al., 2020) datasets, which are common benchmarks for fake news detection tasks. Table 3 shows the statistics of each dataset.

We do not conduct experiments on more datasets due to the long period, as the proposed method involves online data crawling and nested iterative use of LLMs.

4.1.2 Experimental Details

For all datasets, we obtain training, validation, and test sets from the original split data. We use the Scrapy framework written in Python as the tool to crawl the user interaction networks of the news in each dataset. The large language model we use is GPT-4, which is called through the general interface on the official website. For parameter settings, following Devlin et al. (2019), we truncate the input length to 512 and set the vector dimension to 768. The task label classifier adopts a feed-forward neural network with a single hidden layer of 256 neurons. During training, the initial learning rate for the Adam optimizer (Kingma and Ba, 2014) is tuned by grid searches from 1e-6 to 1e-2.

Domain	PolitiFact	GossipCop	CoAID
#Fake News	269	1269	135
#Real News	230	2466	1568

Table 1: The statistics of PolitiFact, GossipCop and CoAID datasets.

4.1.3 Baselines

To demonstrate the effectiveness of the proposed model DUPS, we compare it with several existing methods in three groups:

(1) Early neural network based methods, including:

- **BiGRU** (Ma et al., 2016), is a widely used baseline for fake news detection. We adopt a one-layer BiGRU with a hidden size of 512.
- **BERT** (Devlin et al., 2019), is a pre-training model, which is widely used in various tasks and serves as a commonly baseline.

(2) Traditional fake news detection methods, including:

- **dEFEND** (Shu et al., 2019a), it utilizes sentence-post co-attention network for fake news detection.
- **M^{3}FEND** (Zhu et al., 2022), is a complex fake news detection model, which encodes the news piece from a multi-view perspective and adopts a Memory Bank to enrich information for samples.

(3) Social media fake news detection methods, including:

- **Bi-GCN** (Bian et al., 2020), is a model which can capture the features in terms of the fake news propagation by utilizing bi-directional GCN.
- **WSDMS** (Yang et al., 2023), it only requires bag-level labels for training but is capable of inferring both sentence-level misinformation and article-level veracity.
- **DELL** (Wan et al., 2024), is a method, which use LLM to simulate user-news interaction and generate explanations for each tasks, aiming to enrich the input data.

Model	PolitiFact		GossipCop		CoAID	
	F1	Acc	F1	Acc	F1	Acc
BiGRU	0.572	0.584	0.580	0.569	0.629	0.633
BERT	0.747	0.738	0.713	0.718	0.764	0.755
dEFEND	0.913	0.886	0.756	0.808	0.886	0.870
M^{3}DFEND	0.895	0.877	0.814	0.822	0.911	0.898
Bi-GCN	0.845	0.865	0.805	0.822	0.873	0.857
WSDMS	0.943	0.904	0.870	0.850	0.926	0.893
DELL	0.925	0.906	0.872	0.860	0.881	0.852
DUPS	**0.954**	**0.927**	**0.889**	**0.868**	**0.940**	**0.906**

Table 2: Comparative results on the PolitiFact, GossipCop and CoAID datasets.

4.2 Results

4.2.1 Comparative Experiments

We compare proposed DUPS with seven baselines on three datasets. The main results are shown in Table 2, from which we have the following observations:

(1) Methods based on interaction network modeling generally outperform than other baselines, which reflects that the extraction of relationship features is crucial in social media fake news detection task.

(2) Although these baselines that model complex relational structures achieved impressive results, none of them outperformed others on all three datasets. This suggests that breaking through the performance bottleneck of existing work may require focusing on other aspects, such as data bias.

(3) Compared with all baselines, DUPS achieves the best experimental results on both datasets, which shows the effectiveness of our method.

4.2.2 Ablation Study

In order to verify the effectiveness of the main components in each step of DUPS, we conduct a ablation study for the method on three datasets.

First, we conduct experiments to explore the contributions of five personal attributes, which is used for user portraits analysis in the first step of DUPS. The method that using LLM to analyze user portraits by random attributes (given by the LLM itself), is represented as w/o FPA.

Then, we conduct experiments to explore the contributions of LLM-GAN, which is used for comments optimizing in the second step of DUPS. The method that using LLM to simulate user portraits and generate comments directly, is represented as w/o LLM-GAN.

The last, we conduct experiments to explore the contributions of fusion operation, which is used for

Model	PolitiFact		GossipCop		CoAID	
	F1	Acc	F1	Acc	F1	Acc
DUPS	**0.954**	**0.927**	**0.889**	**0.868**	**0.940**	**0.906**
w/o FPA	0.937	0.916	0.874	0.850	0.921	0.885
w/o LLM-GAN	0.926	0.909	0.867	0.843	0.912	0.872
w/o FO	0.943	0.915	0.880	0.854	0.928	0.890

Table 3: Results of ablation study on the PolitiFact, GossipCop and CoAID datasets.

Metrics	PolitiFact		GossipCop		CoAID	
	Org	De	Org	De	Org	De
avg_num_com	2.1	5.2	2.0	6.3	3.3	7.2
avg_var_com	3.3	0.9	4.5	1.4	4.1	1.9

Table 4: Statistics of the original (Org) and debiased (De) dataset, where avg_num_com represents the average number of comments on each post and avg_var_com represents the average variance of comments on each news.

the combination of real and simulated graph representations in the last step of DUPS. The method that model only the generated debiased data, is represented as w/o FO.

The ablation results are shown in Table 3, from which we have the following conclusions:

(1) All components contribute to the overall performance of the method, which confirms the effectiveness of each step of DUPS.

(2) LLM-GAN contributes the most to the overall performance, which means that LLMs have a room for improvement when doing role-playing.

5 Analysis

In this section, we conduct analysis experiments to answer the following questions:
Q1: How effective is DUPS in debiasing?
Q2: How DUPS improves model predictions on real-world examples?

5.1 Data Analysis (Q1)

We explore the effectiveness of DUPS on data debiasing from two perspectives: data statistics and model performance comparison.

Data statistics: We calculated the average number of comments on each post and the average variance of comments on each news post in the original data and debiased data, as shown in Table 4. From this table, we can conclude that for a given piece of news, using the debiased dataset to model its user interaction information will obtain a larger and more balanced graph network, which can improve the performance of the model.

Model	PolitiFact		GossipCop		CoAID	
	F1	Acc	F1	Acc	F1	Acc
BERT	0.747	0.738	0.713	0.718	0.764	0.755
BERT + debiased	0.770	0.754	0.737	0.733	0.790	0.779
dEFEND	0.913	0.886	0.756	0.808	0.886	0.870
dEFEND + debiased	0.927	0.895	0.768	0.822	0.900	0.887
Bi-GCN	0.845	0.865	0.805	0.822	0.873	0.857
Bi-GCN + debiased	0.860	0.872	0.826	0.836	0.896	0.878

Table 5: Performance comparison of three models on original and debiased data.

model performance comparison: To verify the effectiveness of debiased data in improving model performance, we selected one baseline from each group (which is mentioned in Section 4.1.3) and compared their performance on the original and debiased data. The comparison results are shown in Table 5, from which we can conclude that the three types of baselines perform better on the debiased data, which means that the data is of higher quality for the social media fake news detection task compared to the original one.

In view of the above analysis, the answer to Q1 is: DUPS can obtain high-quality debiased data, which can improve models performance.

5.2 Case Study (Q2)

To explore how DUPS improves model predictions on real-world scenarios, we select an example from twitter, which has been presented in Figure 2. We add it to the test set, then use the baseline model and DUPS to predict its label, respectively.

For this example, we first use the comparison baseline: Bi-GCN, to model its social network. The subgraph weight and final predicted label of news are shown in the left half of Figure 4. Then, we use the proposed DUPS to model this example. The first step is to generate corresponding comments by simulating user portraits; the second step is to model the social network based on the generated debiased data. The corresponding subgraph weight and final predicted label of news are shown in the right half of Figure 4. By comparing the modeling performance and prediction results of the two methods, we have the following conclusions:

(1) In social media scenarios, when performing graph modeling, the model pays more attention to nodes with richer relationships and ignores those important but sparsely related nodes.

(2) The proposed DUPS can alleviate the impact of biased data on models by using LLM for data

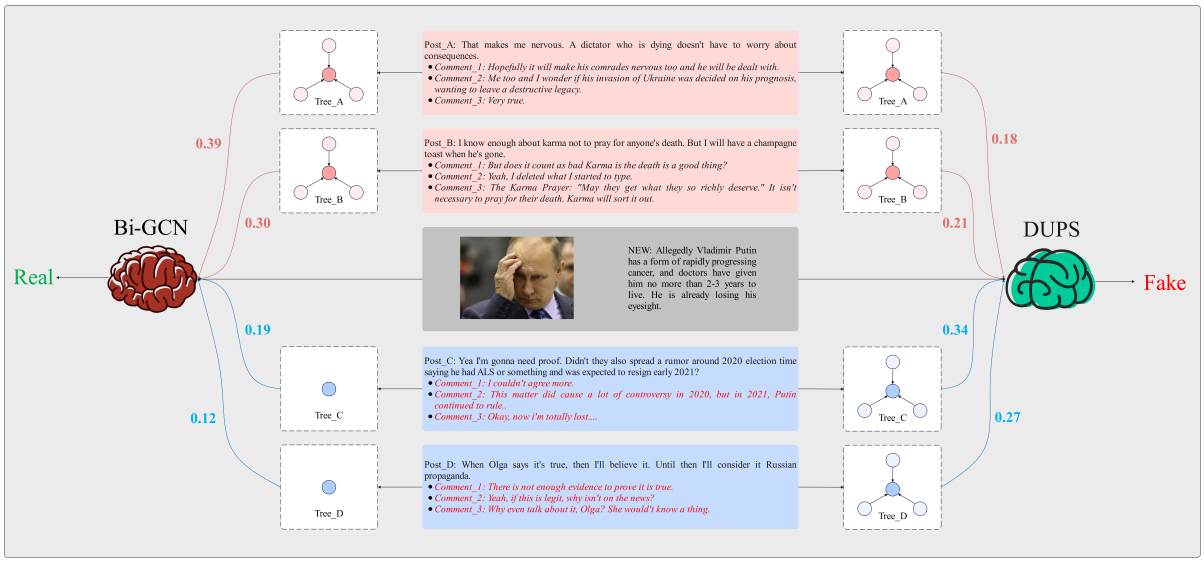
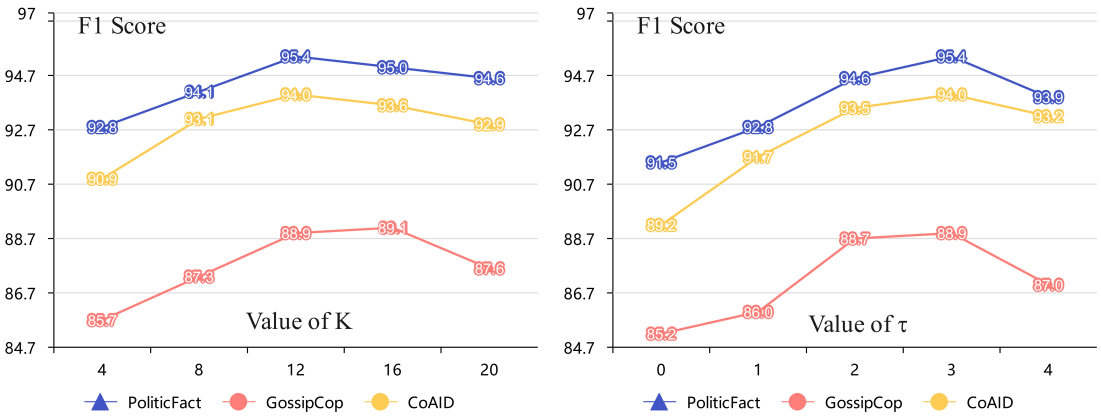


Figure 4: A real-word example from twitter for case study. Red text represents comments generated by simulating user portraits using DUPS.



(a) Effect of the parameter K.

(b) Effect of the parameter τ.

Figure 5: The impact of the number of user portraits clusters (K) and the interest threshold (τ).

augmentation, thereby improving performance.

5.3 Parameter Analysis

In this section, we test the sensitivity of two hyper-parameters used in DUPS: K, which is the number of user portraits clusters mentioned in Section 3.2.2; τ, which is the interest threshold mentioned in Section 3.3.1.

As shown in Figure 5, both parameter K and τ have a certain impact on the model performance, and with properly tuned (K=12 and τ=3), DUPS can achieve satisfying performance. As the parameters increase, the model performance shows a trend of rising and then falling, from which we have the following conclusion:

When using LLMs for role-playing, it is necessary to control the amount and granularity of data generated, otherwise it will be counterproductive.

6 Conclusion

In this paper, we propose a approach that uses LLMs to debias through data augmentation, named DUPS. The method first uses the LLM to analyze the user portraits, and then simulates the corresponding portrait to generate interactive comments, thereby reconstructing unbiased data. At last, the debiased data is modeled to predict news labels. Experimental results on three datasets show that DUPS outperforms the current State-Of-The-Art approaches. In addition, relevant analysis also proves the effectiveness of the method in debiasing, to obtain high-quality debiased data.

Limitations

This work has two limitations: The experimental dataset is not sufficient and the generated content may not be directly used to detect the authenticity of news. For the first limitation, the reason is that the experimental period is too long, which has been mentioned in Section 4.1.1. After improving the time consumption of this work, we will conduct it on more datasets. As for the second limitation, it is a common problem in role-playing methods, due to the enhanced data may not necessarily be used as the basis for task label prediction. In future work, we consider adopting the RAG (Retrieval Augmented Generation) approach to ensure the reliability of the generated data.

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Lucas Azevedo, Mathieu d'Aquin, Brian Davis, and Manel Zarrouk. 2021. LUX (linguistic aspects under eXamination): Discourse analysis for automatic fake news classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 41–56, Online. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? *Preprint*, arXiv:2309.13788.

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.

Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User preference-aware fake news detection. *Preprint*, arXiv:2104.12259.

Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 877–880, New York, NY, USA. Association for Computing Machinery.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *ArXiv*, abs/1703.09398.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024. Large language model agent for fake news detection. *Preprint*, arXiv:2405.01593.

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

Annabelle GY Lim. 2023. Big five personality traits: The 5-factor model of personality. *Simply Psychology*.

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 (EMNLP-IJCNLP)*, pages 4821–4830, Hong Kong, China. Association for Computational Linguistics.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard Jim Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *International Joint Conference on Artificial Intelligence*.

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.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *Preprint*, arXiv:2310.10158.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019b. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *Preprint*, arXiv:1809.01286.

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 Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 51–59, New York, NY, USA. Association for Computing Machinery.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *Preprint*, arXiv:2402.10426.

Yifeng Wang, Zhouhong Gu, Siwei Zhang, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2024. Llm-gan: Construct generative adversarial network through large language models for explainable fake news detection. *Preprint*, arXiv:2409.01787.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. *Preprint*, arXiv:2402.09269.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Zhiwei Yang. 2023. WSDMS: Debunk fake news via weakly supervised detection of misinforming sentences with contextualized social wisdom. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1525–1538, Singapore. Association for Computational Linguistics.

Yuting Yang, Juan Cao, Mingyan Lu, Jintao Li, and Chia-Wen Lin. 2021. How to write high-quality news on social network? predicting news quality by mining writing style. *Preprint*, arXiv:1902.00750.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *Preprint*, arXiv:1809.05679.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. *Preprint*, arXiv:1909.04465.

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.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. *Preprint*, arXiv:2305.14871.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, page 1–14.

A Example Appendix