Equal Truth: Rumor Detection with Invariant Group Fairness

Junyi Chen[1], Mengjia Wu[2], Qian Liu[1], Jing Sun[1], Ying Ding[3], Yi Zhang[2]*
[1]University of Auckland [2]University of Technology Sydney [3]University of Texas at Austin
{junyi.chen, liu.qian, jing.sun}@auckland.ac.nz,
{mengjia.wu, yi.zhang}@uts.edu.au,
ying.ding@ischool.utexas.edu

Abstract

Due to the widespread dissemination of rumors on social media platforms, detecting rumors has been a long-standing concern for various communities. However, existing rumor detection methods rarely consider the fairness issue inherent in the model, which can lead to biased predictions across different stakeholder groups (e.g., domains and originating platforms of the detected content), also undermining their detection effectiveness. In this work, we propose a two-step framework to address this issue. First, we perform unsupervised partitioning to dynamically identify potential unfair data patterns without requiring sensitive attribute annotations. Then, we apply invariant learning to these partitions to extract fair and informative feature representations that enhance rumor detection. Extensive experiments show that our method outperforms strong baselines regarding detection and fairness performance, and also demonstrate robust performance on out-of-distribution samples. Further empirical results indicate that our learned features remain informative and fair across stakeholder groups and can correct errors when applied to existing baselines.

1 Introduction

Social media has reshaped the convenience of how people exchange their daily information, but it has also facilitated the spread of rumors via the internet. A rumor, defined as a piece of fabricated information, aims to mislead the public and generate illegal profits. Therefore, detecting rumors accurately and promptly has become a shared goal in society.

Previously, most rumor detectors (Devlin et al., 2019; Nan et al., 2021; Wang et al., 2018; Azri et al., 2021; Zhang et al., 2021) utilize data-driven methods to learn news content representations for rumor classification. This follows an ideal causal pathway $x \rightarrow y$, where x denotes news content and y represents the classification label. However, they often neglect the unseen confounding effect modeled as $x \leftarrow s \rightarrow y$, where a sensitive attribute s indicates stakeholder group membership (e.g., domain) of the content. Taking domain-based groups as an example, $x \leftarrow s$ manifests as domain-specific linguistic patterns shape the content (e.g., science versus politics). The path $s \rightarrow y$ manifests through data collection biases, i.e., the disparate class distribution across domains (Zhou et al., 2024; Nan et al., 2021; Li et al., 2024c). This extends to other sensitive attributes (e.g., platform), where news from major agencies (e.g., CNN news) differs linguistically from social media streams (e.g., Twitter/Reddit) and is stereotyped as more trustworthy. We derive empirical findings in Fig. 1 to support the above causal analysis. It is evident that non-causal shortcuts can be learned due to discrepancies in feature and class distributions across groups. As shortcuts allow the model to infer group identity and characteristics of the content, data-driven rumor detection methods are prone to making biased and unfair decisions, often classifying content from high-rumor-ratio groups as rumors without rigorously verifying authenticity.

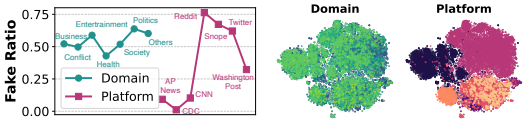


Figure 1: Preliminary findings on a recent benchmark (Zhou et al., 2024). Left: rumor ratio (y-axis) across domains and platforms (x-axis). Right: t-SNE visualization of content features obtained by BERT (Devlin et al., 2019), illustrating clear distinctions across domains and platforms.

Current studies in rumor detection have only examined the bias issue in training datasets. Specifically, while certain types of bias (e.g., entity (Zhu et al., 2022a) and psycholinguistic bias (Chen et al., 2023)) have been partially addressed, these mitigation approaches can hardly be quantitatively mea-

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10994–11007

sured in terms of fairness for stakeholders from diverse groups (e.g., across different content domains and platforms). To enhance rumor detection from a group fairness perspective, we identify two critical challenges. **1. Limited annotated but diverse sensitive attributes.** Privacy constraints and massive media data prevent comprehensive sensitive attribute annotation related to the detected content (e.g., domains, platforms, authors' certification status, political leaning, etc.), yet each attribute defines its own set of groups. **2. Variant feature learning.** Limited sensitive attribute supervision produces features that remain unfairly variant to unknown sets of groups, compromising generalization of fairness and detection effectiveness.

Ideally, the model should accurately detect rumors while avoiding disproportionate predictions within each identifiable set of stakeholder groups to ensure fairness. To approach this, we develop FIRM, which stands for Fair and Invariant Rumor detection Model for multiple sets of groups. This majorly takes two steps. Initially, we split the training data into different subsets using a parameterized neural network to identify a potential unfair data partition. Subsequently, we leverage invariant learning to improve and balance the model's performance across each subset within the partition. The above two challenges are effectively addressed by our method: 1. it does not require any sensitive attribute annotations to perform data partitioning, and 2. it can account for a wide range of stakeholder groups, as the discovered unfair partition dynamically evolves with the framework's parameter updates during training. Hence, the learned feature representations should be invariant and generalizable across diverse stakeholder groups.

During experimental evaluation, we assess both detection effectiveness—using metrics such as accuracy and F1 score—and fairness, measured by maximum demographic disparity (Barocas et al., 2023), which captures whether predictions are disproportionately distributed within each set of group defined by ground-truth sensitive attributes. Results demonstrate that our method performs strongly on both fronts compared to strong baselines. Our key contributions are summarized as follows:

- We study rumor detection from a fairness perspective regarding multiple sets of groups, a novel angle largely overlooked in previous research.

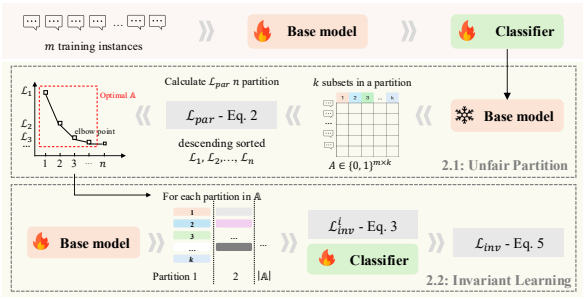


Figure 2: Overview of our proposed method.

- We propose a unified fairness framework that works for various sets of groups without requiring sensitive attribute annotations.

- Extensive experiments show that our method consistently surpasses state-of-the-art rumor detection baselines in both effectiveness and fairness, while also exhibiting superior robustness and plug-and-play adaptability.

It is important to acknowledge that while fairness considerations and sensitive attributes were traditionally associated with legally protected characteristics (Chen et al., 2019; Liu et al., 2020) and demographic categories (e.g., gender, age, and race), recent studies in online social application systems (e.g., recommender systems (Fu et al., 2020; Naghiaei et al., 2022; Wang et al., 2022)) have expanded these concepts to encompass stakeholders from diverse groups, which constitute a significant component of online social information communities. Hence, in this study, we also loosely use the term "sensitive attributes", based on which we can link to different stakeholder groups relevant to the benefits of the rumor detector's outcome.

2 Method

Problem Formulation Let $c = \{\boldsymbol{x}, y, \boldsymbol{S}\}$ represent a rumor detection instance, where $\boldsymbol{x}$ denotes the initial embedded feature, y denotes the classification label, and $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ denotes the sensitive attribute vectors indicating group membership of different sets (e.g., $\boldsymbol{s}_1$ indicates the domain, $\boldsymbol{s}_2$ indicates the platform). Given a rumor detector with a base model $\Phi(\cdot)$[1] for feature extraction and a classifier $f(\cdot)$, we aim to optimize both for accurate binary prediction $y \in \{\text{rumor}, \text{non-rumor}\}$ while ensuring fairness across groups from all po-

[1]Unless otherwise specified, all experiments in this study utilize the 'bert-base' (Devlin et al., 2019) as our base model.

tential sets. $\boldsymbol{S}$ serves only for fairness evaluation. We omit the subscript when the context is clear.

Overview The overall framework of our method is shown in Fig. 2. Our method starts with standard training, then iteratively performs two steps: (1) *Unfair Partition*, which discovers data partitions that degrade the performance of the base model, and (2) *Invariant Learning*, which aims to improve and balance performance in these partitions.

2.1 Unfair Partition

The concept of backdoor adjustment (Pearl et al., 2016; Zhang et al., 2022) suggests partitioning data by sensitive attributes to identify unfair distributions for learning fair representations. However, this approach demands comprehensive attribute annotations - impractical for social media data and risks overlooking unidentified sets of groups.

Inspired by invariant learning (Li et al., 2024a; Zhu et al., 2024; Arjovsky et al., 2019), we propose dynamic data partitioning to consider multiple sets of groups during training, without requiring sensitive attribute annotations. Our method identifies "unfair partitions" where the frozen base model performs worst. Recalling the ideal causal pathway $\boldsymbol{x} \rightarrow y$, if the base model learns to extract causal features and is unaffected by confounding sensitive attributes, it should generally produce well-separated features based on contrastive labels y within each subset given a partition. Otherwise, it may be influenced by confounding attributes via the pathway $\boldsymbol{x} \leftarrow s \rightarrow y$, where the confounding effect of s varies across instances within and between subsets, leading to non-causal, non-generalizable features and thus, the worst performance within a partition. In this context, subsets in the partition can be interpreted as a set of groups conditioned on the specific type of confounding sensitive attribute (e.g., content from the domain of science versus politics).

Based on the discussion above, for an instance c with its initial embedded feature $\boldsymbol{x}$, we first deduce the partition:

$$\hat{\boldsymbol{s}} = FC(\Phi(\boldsymbol{x})), \tag{1}$$

where $FC(\cdot)$ is a fully-connected layer with trainable parameters and $\hat{\boldsymbol{s}} \in \mathbb{R}^{1\times k}$ indicates which subset the instance belongs to and k is a hyperparameter that defines the number of the split subsets.

By applying this operation across m training instances, we construct a partition matrix $\boldsymbol{A} \in \{0,1\}^{m\times k}$, which we want to potentially reflect data distribution under a specific set of groups. $\boldsymbol{A}$ is parameterized by $FC(\cdot)$, enabling backpropagation. Based on the objective of invariant risk minimization (Arjovsky et al., 2019), we identify unfair partitions where the base model performs worst, helping reveal confounding sensitive attributes:

$$\mathcal{L}_{\text{par}} = \underset{\boldsymbol{A}}{\text{argmax}} \sum_{k} \mathcal{L}_{\text{con}}(\boldsymbol{A}, k, \Phi(\boldsymbol{x}), y) + \lambda \,\text{Var}\left(\{\mathcal{L}_{\text{con}}(\cdot)\}_{i=1}^{k}\right), \tag{2}$$

$$\mathcal{L}_{con} = \sum_{(x,x^+)\in \boldsymbol{A}_{\{*,k\}}} -\log \frac{e^{\Phi(x)^\top \Phi(x^+)}}{\sum e^{\Phi(x)^\top \Phi(x^*)}} \tag{3}$$

where $\mathcal{L}_{\text{con}}$ denotes the supervised contrastive loss for each subset given a partition, $\text{Var}(\cdot)$ measures performance variance across subsets and λ is a trade-off factor. In implementation, we treat the items within the same subset that share the same label y as positive pairs, while the remaining items within the same subset are treated as negative pairs. A higher value of $\mathcal{L}_{\text{con}}$ indicates the base model's reliance on a potential confounding sensitive attribute - a non-causal correlation that limits detection performance. A higher value of $\text{Var}(\cdot)$ indicates performance disparity across subsets. Using supervised contrastive loss directly help us measure the feature variance and thus reveal potential confounding sensitive attributes.

At this stage, we freeze the base model. Each epoch's partition matrix $\boldsymbol{A}$ is added to set $\mathbb{A}$, enabling continuous discovering of unfair partitions caused by different confounding attributes and thus, different sets of groups. However, this approach introduces substantial computational complexity in the second step, and not all partitions contribute equally to improving detection performance and fairness. To address this, we implement a parameter-free elbow point detection method derived from information theory (Baptista et al., 2021; Antunes et al., 2018) to adaptively identify distinct data patterns—specifically, to determine which unfair partitions are both meaningful and distinct. Given n partitions, we proceed as follows:

1. Sort the partition losses $\mathcal{L}_{\text{par}}$ in descending order to obtain $L_1, L_2, ..., L_n$ where $L_1 \geq L_2 \geq ... \geq L_n$, with corresponding ordered partitions $\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_n \in \mathbb{A}$, and compute normalized losses $\tilde{L}_i = \frac{L_i}{L_1}$.

2. Calculate the first and second-order differences: $\Delta_i = \tilde{L}_i - \tilde{L}_{i+1}$ for $i \in \{1, 2, \ldots, n-1\}$ and $\kappa_i = \Delta_i - \Delta_{i+1}$ for $i \in \{1, 2, \ldots, n-2\}$.

3. Identify the elbow point $i^* = \underset{i \in \{1,2,\ldots,n-2\}}{\operatorname{argmax}} \kappa_i$ and select the optimal partition set $\mathbb{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{i^*+1}\}$.

The elbow point serves as a threshold indicating where additional partitions cease to provide significant new information.

2.2 Invariant Learning

The invariant learning step eliminates non-causal correlations potentially induced by confounding sensitive attributes within each partition. On one hand, it drives both the base model and classifier to improve detection performance, and on the other hand, it promotes balanced performance across groups for $\forall \boldsymbol{A} \in \mathbb{A}$, to learn more invariant features. Hence, for a specific partition, the goal in the second step is to flip the objective to minimize it:

$$\begin{aligned} \mathcal{L}_{\text{inv}}^{i} = \min_{\Phi, f} \sum_{k} \mathcal{L}_{\text{cls}}(\boldsymbol{A}_i, k, f, \Phi(\boldsymbol{x}), y) \\ + \lambda \operatorname{Var}\left(\{\mathcal{L}_{\text{cls}}(\cdot)\}_{i=1}^{k}\right), \boldsymbol{A}_i \in \mathbb{A} \end{aligned} \tag{4}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss used for classification. However, treating all partitions with equal weight is suboptimal, as this approach fails to reflect real-world scenarios where certain group partitions may experience more severe fairness issues and consequently constrain detection performance more significantly. To address this limitation, we propose calculating an importance score for each partition $\boldsymbol{A}_*$. The importance score is formulated as:

$$\alpha_i = \frac{e^{\mathcal{L}_{\text{inv}}^{i}}}{\sum_{j=1}^{|\mathbb{A}|} e^{\mathcal{L}_{\text{inv}}^{j}}} \tag{5}$$

where $\mathcal{L}_{\text{inv}}^{i}$ represents the invariant loss corresponding to partition i. This formulation enables our method to allocate greater attention to partitions exhibiting higher loss values, which is captured by the following weighted objective function:

$$\mathcal{L}_{\text{inv}} = \sum_{i=1}^{|\mathbb{A}|} \alpha_i \mathcal{L}_{\text{inv}}^{i} \tag{6}$$

	Chinese	*English*
Total samples	23,969	16,909
Rumor	17,895 (74.7%)	9,407 (55.6%)
Non-rumor	6,074 (25.3%)	7,502 (44.4%)
# Domains	12	7
# Platforms	17	7
# Author cerification status	2	2

Table 1: Statistics of rumor detection datasets.

3 Experimental Evaluation

3.1 Setup

Datasets. We conduct experiments using four public datasets spanning English and Chinese texts, including: 1. *FineFake* (Zhou et al., 2024), 2. *Weibo21* (Nan et al., 2021), and 3. *MCFEND* (Li et al., 2024c). To facilitate presentation and demonstrate the group fairness and detection effectiveness we aim to improve, following common practices in existing literature (Zhu et al., 2022a; Bu et al., 2024; Li et al., 2024b), we refer the first dataset as the *English* dataset and the latter two as the *Chinese* dataset. Both datasets include three sensitive attributes defining distinct groups: Domain (7 in English, 12 in Chinese), Platform (7 in English, 17 in Chinese), and Author Certification Status (binary in both datasets). We've summarized the statistical distribution of the datasets in Table 1, and a more detailed version in Appendix A.4.

Baselines. We select a set of state-of-the-art baselines for comparison, categorized as follows: 1. Purely content-based methods: BERT; 2. Multi-domain methods: EANN (Wang et al., 2018), MDFEND (Nan et al., 2021), and M3FEND (Zhu et al., 2022b); 3. Comment-based methods: DualEmo (Zhang et al., 2021), dFEND (Shu et al., 2019); 4. Debiasing methods: ENDEF (Zhu et al., 2022a), CDD (Chen et al., 2023), DTDBD (Li et al., 2024b); 5. Large Language Model (LLM)-based methods: Fine-tuned DeepSeek (Liu et al., 2024) and GPT-4o (Achiam et al., 2023) on respective datasets. Detailed descriptions and implementations are in Appendix A.1.

Evaluation Protocol. We split all datasets into training, validation, and test sets with ratios of 3:1:1 and remove all duplicate entries to prevent data leakage. The best checkpoint on the validation set is used for testing, and results are averaged over 10 runs. For evaluation, we use accuracy (Acc.) and F1 score to measure detection effectiveness, and Maximum Demographic Parity (Barocas et al., 2023), $\Delta = \max_{s,s'} \left| P(y = 1 \mid s) - P(y = \right.$

Model	*English*					*Chinese*				
	Acc ↑	F1 ↑	$\Delta_d \downarrow$	$\Delta_p \downarrow$	$\Delta_a \downarrow$	Acc ↑	F1 ↑	$\Delta_d \downarrow$	$\Delta_p \downarrow$	$\Delta_a \downarrow$
BERT	77.1	76.9	42.3	86.7	37.2	82.3	76.9	55.6	49.9	18.3
EANN	77.2	76.4	34.2	88.6	41.4	83.2	77.3	**40.3**	54.1	15.4
MDFEND	77.9	77.6	44.1	92.1	34.6	<u>84.2</u>	<u>78.2</u>	58.3	53.6	14.2
M3FEND	78.1	77.7	44.4	86.4	37.3	84.0	77.5	60.4	51.3	16.3
DualEmo	78.5	78.0	39.2	86.9	30.3	83.9	76.5	54.3	56.8	17.9
dFEND	77.9	76.6	40.2	85.4	37.8	82.1	76.2	53.6	52.3	15.4
ENDEF	77.1	76.2	34.9	88.6	38.1	83.4	77.7	48.6	55.8	<u>13.9</u>
CDD	78.5	<u>78.1</u>	35.3	91.6	35.5	83.8	77.3	51.1	53.4	14.8
DTDBD	<u>79.6</u>	77.9	<u>33.1</u>	89.2	36.3	82.6	76.8	49.8	56.2	16.3
DeepSeek	78.2	77.8	39.7	82.1	30.1	80.5	73.4	50.2	<u>51.3</u>	14.6
GPT-4o	77.7	77.0	43.3	<u>79.6</u>	<u>29.8</u>	82.1	75.6	51.9	53.1	18.6
FIRM (Ours)	**81.1**	**80.5**	**27.3**	**77.3**	**27.2**	**88.2**	**82.9**	<u>45.5</u>	**46.6**	**11.3**

Table 2: Performance comparison. The best value is in **bold** while the runner-up is <u>underlined</u>

$1 \mid s')\big|$, where (s, s') denotes any pair of distinct groups within a set, to assess algorithmic fairness. The metric $\Delta_{d/p/a}$ denotes this measure within groups defined by domains, platforms, and authors. We set $k = \{2, 3\}$ and $\lambda = \{0.6, 5\}$ for the English and Chinese datasets respectively, supported by observations from the hyperparameter analysis in Section 3.8.

3.2 Performance Comparison

We provide the results in Table 2, which summarizes key insights: 1. BERT, purely content-based, performs worst in detection and fairness, reflecting its difficulty in handling diverse sources and bias. 2. Multi-domain methods improve detection but increase unfairness, as domain-specific information may reinforce stereotypes. 3. Debiasing methods often improve detection and domain-specific fairness but may degrade fairness across other group sets, likely due to their focus on domain-related biases. 4. Comment-based approaches enhance detection but introduce unfairness, likely due to differing user interaction habits. 5. LLM-based methods show promise but still face fairness issues, despite strong detection performance. 6. Our method outperforms others in both detection and fairness, thanks to a two-step process that identifies and optimizes unfair partitions, promoting diversity and generalizability.

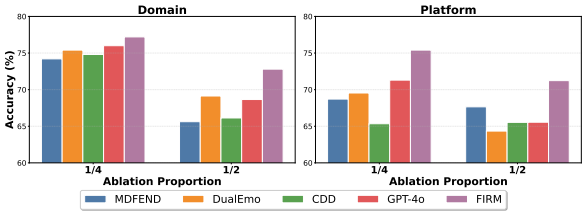
3.3 Ablative Study

We conduct a comprehensive ablative study to verify our design motivations, deriving the following variants: (1) w/ SP: using static partitioning on training data with all ground truth sensitive attribute labels instead of our dynamic strategy; (2) w/o $\mathcal{L}_{\text{sup}}$: replacing the supervised contrastive loss with cross-entropy in the unfair partition step; (3) w/o Var: removing the variance loss between partitioned subsets for both stages; (4) w/o elbow: without using the elbow point detection to determine partitions, instead fixing four partitions with the highest loss, an optimal choice based on hyperparameter analysis; (5) w/o imp: removing the importance-weighted loss, treating all partitions equally; (6) w/o Record: not recording every partition during training, only using the latest one.

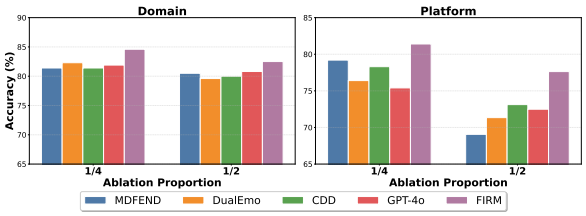
Results in Table 3 reveal that w/ SP achieves comparable fairness but poorer detection, as static partitioning struggles to find informative invariant features due to confounding effects that are highly elusive in rumor detection. Replacing $\mathcal{L}_{\text{sup}}$ with cross-entropy results in worse performance, indicating supervised contrastive loss better captures variant feature patterns caused by sensitive attributes. Dropping the variance loss hampers learning of generalizable features. Using a fixed number of high-loss partitions is suboptimal compared to adaptive elbow point detection, which avoids redundancy and finds more informative patterns across distributions. The importance score ensures that severely impaired partitions receive more focus. Finally, the w/o Record variant shows that optimizing all discovered partitions throughout training, rather than only the latest, is essential for robustness.

Variants	*English*		*Chinese*	
	Acc	$\Delta_{d/p/a} \downarrow$	Acc	$\Delta_{d/p/a} \downarrow$
FIRM	81.1	27.3/77.3/27.2	88.2	45.5/46.6/11.3
w/ SP	78.8	25.8/78.6/25.2	84.9	44.2/48.8/13.3
w/o $\mathcal{L}_{\text{sup}}$	79.4	27.5/81.4/33.7	85.4	45.7/48.5/13.9
w/o Variance	78.7	31.2/85.0/39.5	83.6	51.9/56.9/17.1
w/o elbow	79.3	29.2/81.6/28.5	86.1	47.0/47.9/13.1
w/o imp	79.3	28.1/80.6/29.4	86.8	46.5/50.9/13.8
w/o Record	78.5	23.3/86.3/33.7	86.4	46.5/54.4/15.2

Table 3: Ablative study on model variants.



(a) English dataset



(b) Chinese dataset

Figure 3: OOD test results for baselines and our method on English and Chinese datasets.

3.4 Robustness to OOD samples

In real-world scenarios, collecting comprehensive training data is impractical, making out-of-distribution (OOD) samples inevitable post-deployment. To evaluate robustness, we conduct a group category ablation experiment across three groups: domains, platforms, and author certification status. For each group, we remove a percentage of categories from the training set, leaving them only in the test set, covering all possible ablations to report mean performance. Since the author group has only two categories, we do not perform experiments on it due to data sparsity.

Results on OOD test sets in Fig. 3 show that all methods perform worse than in the normal setting, highlighting the importance of group diversity for robustness. Ablating platform categories causes the most significant performance decline, indicating that platform-specific content contains valuable patterns for rumor detection and that models trained on a single platform are vulnerable in real-world scenarios. For domain ablation, the performance drop in multi-domain and debiasing methods is more pronounced, reflecting their reliance on domain information. Our method remains more robust than baselines, thanks to its automatic partition strategy, which does not depend on ground-truth labels and can discover multiple data patterns, helping the model learn more general, invariant features and improving OOD robustness. We also provide more detailed performance results, including mean and standard deviation values, in table form in Appendix A.3.

Baselines	*English*		*Chinese*	
	Acc	$\Delta_{d/p/a} \downarrow$	Acc	$\Delta_{d/p/a} \downarrow$
EANN	77.2	34.2/88.6/41.4	83.2	40.3/54.1/15.4
w/ours	80.5	31.3/82.3/30.4	87.9	40.5/45.1/10.1
MDFEND	77.9	44.1/92.1/34.6	84.2	58.3/53.6/14.2
w/ours	81.9	30.1/80.6/24.3	89.1	49.7/43.4/12.5
M3FEND	78.1	44.4/86.4/37.3	84.0	60.4/51.3/16.3
w/ours	81.5	33.7/79.6/29.8	88.9	50.2/46.7/12.5

Table 4: Baseline integration study.

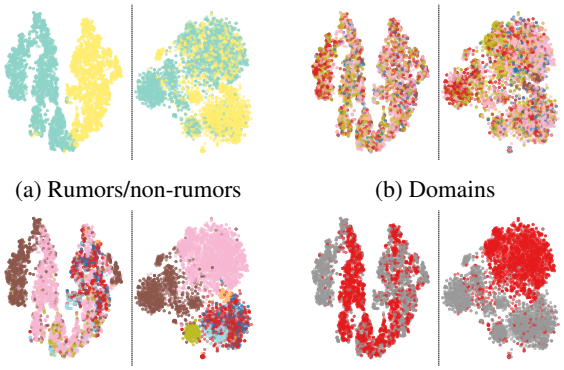
3.5 Baseline Integration and Enhancement

The proposed framework is flexible and can be adapted to different baselines by changing the feature extractor. In the above experiments, we only used BERT. In this experiment, we integrate our framework with existing state-of-the-art baselines. Since our method is primarily designed for content-based rumor detection, and the multi-domain-based methods have demonstrated strong performance, we incorporate our method into these baselines.

We present the experimental results in Table 4. We observe that: 1. By integrating our proposed framework, the detection effectiveness and fairness of all baseline methods are improved. This indicates that the proposed method is a general framework that can readily refine the predictions of content-based rumor detection models. 2. More specifically, with stronger baseline models (as reflected in the overall performance comparison), the improvements become more noticeable and significant.

3.6 Visualization and Intervention

To further demonstrate that our method enhances rumor detection while mitigating unfairness from confounding sensitive attributes, we visualize learned features from the multi-domain-based method MDFEND and our approach integrated with it. Points are color-coded by ground-truth rumor labels, domain, platform, and author certification status on the English test set. Visualizations



(a) Rumors/non-rumors (b) Domains

(c) Platforms (d) Authors

Figure 4: Visualization of learned features. Left: MDFEND w/ our method. Right: MDFEND. Colors denote different groups.

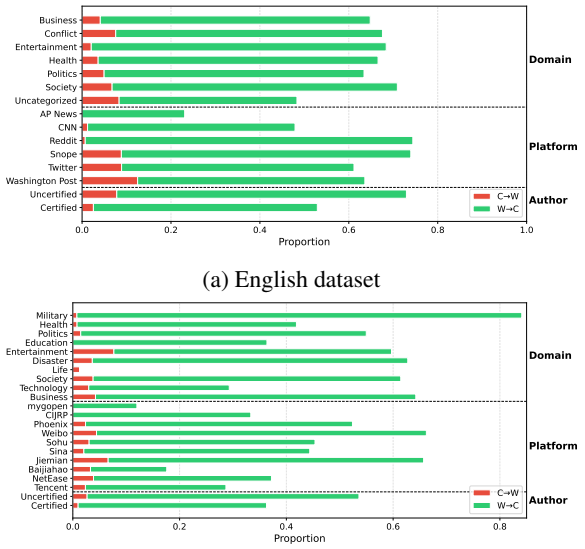
on the Chinese test set, with consistent observations, are provided in Appendix A.2.

As shown in Fig. 4, the learned representations from our method, compared to MDFEND, exhibit: (1) stronger discriminative power for distinguishing rumors from non-rumors, and (2) reduced vulnerability to confounding sensitive attributes. In MDFEND, points often cluster by group, with scattered inter-group distributions that encourage shortcut learning. In contrast, our method produces uniformly mixed and tightly clustered representations across groups, with diverse feature patterns even for the same group, mitigating bias. Consistent observations across subfigures confirm our method's effectiveness in addressing fairness across different group sets during training.

To examine how our method conducts effective intervention on the base model regarding different groups, we provide further evidence in Fig. 5. It is evident that in most cases where the base model makes incorrect predictions, our method helps correct them, while there is minimal chance that our method will mislead the base model's already correct predictions. We further visualize the distribution of correctly rectified rumors and non-rumors in Fig. 6, showing that our method achieves balanced corrections on the English dataset and primarily corrects false rumors on the Chinese dataset, thereby mitigating over-policing and enhancing fairness. Notably, the correction patterns align with the original rumor/non-rumor distributions in the English (balanced) and Chinese (rumor-dominant) datasets.

3.7 Case Study

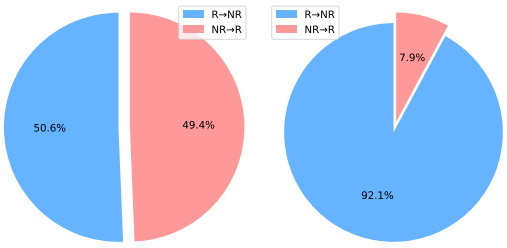
We conduct a case study to demonstrate our model's effectiveness in real-world scenarios. We



(a) English dataset

(b) Chinese dataset

Figure 5: Intervention effects of our method on the base model, MDFEND. W->C: proportion of base model's wrong predictions corrected by our method; C->W: proportion of correct predictions made incorrect.



(a) English dataset (b) Chinese dataset

Figure 6: Corrections distribution after integrating our method into MDFEND. R: Rumor; NR: Non-rumor.

demonstrate one challenging sample from each dataset, where the sample's belonging group is characterized by a high rumor ratio. As shown in Fig. 7, multi-domain-based methods often make incorrect judgments due to learned biases, and current debiasing methods also struggle when multiple confounding sensitive attributes are present. In contrast, our approach effectively removes such biases by learning invariant features across partitions, promoting fairer decisions.

3.8 Hyperparameter and Efficiency Analysis

We investigate the influence of different λ and k on the model's performance in Fig. 8. The results indicate that, across a wide range of λ and k combinations, our method remains generally stable. Increasing k and the trade-off factor λ up to certain thresholds enhances both detection and fairness performance. However, further increases in these

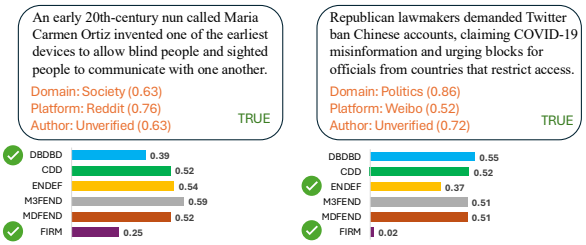
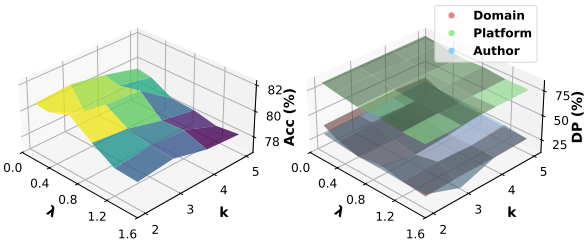
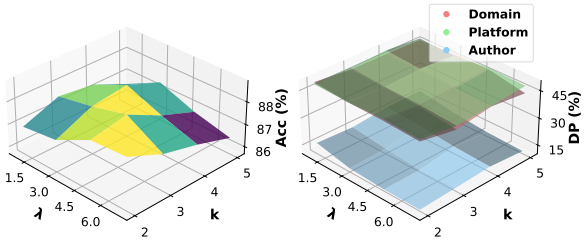


Figure 7: Case study on challenge samples: Left — English dataset; Right — Chinese dataset (translated). The numbers in parentheses denote the rumor ratio for each group of the selected sample. The classification threshold is 0.5, and correct classifications are highlighted.



(a) English dataset



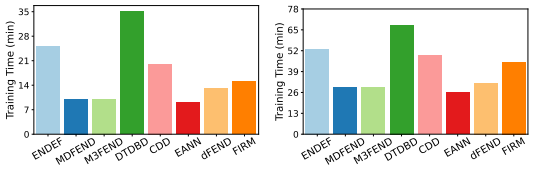
(b) Chinese dataset

Figure 8: Hyperparameter analysis. In each sub-figure, the left shows accuracy variation, and the right displays maximum demographic parity (DP↓) with respect to different combinations of λ and k.

parameters lead to a decline in detection accuracy, as the model sacrifices classification performance to achieve more fair decisions between partitions. We also compare the training efficiency in Fig. 9, which demonstrate our method is generally training efficient.

4 Related Work

Our work is closely related to content-based rumor detection. Earlier research in this domain employed hand-crafted features with traditional machine learning models (Castillo et al., 2011; Kwon et al., 2013) to determine content authenticity. With the advancement of deep learning, most studies have shifted toward pre-trained language models, such as BERT (Devlin et al., 2019) and TextCNN (Zhang and Wallace, 2017), to learn content feature representations. Furthermore, following the success of LLMs (Liu et al., 2024; Achiam et al., 2023), several works (Chen et al., 2025; Hu et al., 2024; Nan et al., 2024; Wang et al., 2024) have leveraged them as feature enhancers or reasoners to augment detection effectiveness. Given that news content frequently emerges from diverse domains, multi-domain approaches (Wang et al., 2018; Nan et al., 2021; Zhu et al., 2022b) have become a prevalent paradigm in this area, employing techniques such as adversarial training, mixture of experts, and memory banks. Additionally, similar to many deep learning-based models, rumor detection systems remain vulnerable to biases inherent in training data. Consequently, some debiasing methods (Chen et al., 2023; Zhang et al., 2022; Zhu et al., 2022a; Li et al., 2024b) have been proposed to mitigate entity bias, psychological bias, and domain bias in the content. However, the biases they address cannot be quantifiably linked to fairness regarding stakeholders from different groups, resulting in limited studies addressing rumor detection from a fairness perspective. In this work, we adopt concepts from invariant learning (Wang et al., 2022; Arjovsky et al., 2019; Zhu et al., 2024; Li et al., 2024a), which is originally designed for OOD generalization—to examine rumor detection from a multi-group fairness perspective.



(a) English dataset (b) Chinese dataset

Figure 9: Training efficiency analysis.

5 Conclusion

In this work, we propose FIRM, which addresses the task of rumor detection from a multi-group fairness perspective, supported by extensive experiments that yield insightful findings. We specifically summarize that: (1) current rumor detection methods exhibit inferior detection effectiveness and fairness across different stakeholder groups, primarily due to confounding sensitive attributes present in real-world data; (2) the confounding sensitive attributes are not limited to a single type, making it important to consider the fairness of multiple group sets during training; (3) jointly leveraging unsupervised partition learning and invariant learning benefits the debiasing of potential confounding

attributes, thereby enhancing both detection effectiveness and fairness.

Limitation

While this study proposes an effective group fairness framework for rumor detection, it has two limitations. First, concerning the dataset used, although we define contextually sensitive attributes, these attributes are not protected by laws (which are not available and defined in current widely used rumor/fake news detection datasets), despite they are still important for developing fair and ethical deep learning algorithms. To broaden the scope and utility of this research, future work could consider developing a more comprehensive fair rumor detection benchmark dataset that includes annotated legally protected attributes for fairness evaluation and yields more insightful empirical findings related to rumor detection on social media platforms. Second, regarding the proposed methodology, we assume that there are no available sensitive attribute annotations for supervised training, which is a very strict condition in real-world scenarios. Future research could explore incorporating sparse supervision signals, where sensitive attribute annotations are used infrequently to perform weakly supervised training, potentially leading to improved empirical performance.

Acknowledgement

This work was supported by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, in conjunction with the National Science Foundation (NSF) of the United States, under CSIRO-NSF #2303037.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mário Antunes, Diogo Gomes, and Rui L Aguiar. 2018. Knee/elbow estimation based on first derivative threshold. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 237–240. IEEE.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Abderrazek Azri, Cécile Favre, Nouria Harbi, Jérôme Darmont, and Camille Noûs. 2021. Calling to cnn-lstm for rumor detection: A deep multi-channel model for message veracity classification in microblogs. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part V 21*, pages 497–513. Springer.

Marcia L Baptista, Elsa MP Henriques, and Kai Goebel. 2021. More effective prognostics with elbow point detection and deep learning. *Mechanical systems and signal processing*, 146:106987.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348.

Junyi Chen, Leyuan Liu, and Fan Zhou. 2025. Do not wait: Preemptive rumor detection with cooperative llms and accessible social context. *Information Processing & Management*, 62(3):103995.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, Minneapolis, MN. Association for Computational Linguistics.

Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 69–78.

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 AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

Dong Li, Chen Zhao, Minglai Shao, and Wenjun Wang. 2024a. Learning fair invariant representations under covariate and correlation shifts simultaneously. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1174–1183.

Jiayang Li, Xuan Feng, Tianlong Gu, and Liang Chang. 2024b. Dual-teacher de-biasing distillation framework for multi-domain fake news detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 3627–3639. IEEE.

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024c. Mcfend: A multi-source benchmark dataset for chinese fake news detection. In *Proceedings of the ACM Web Conference 2024*, pages 4018–4027.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2020. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pages 155–167. Springer.

Mohammadmehdi Naghiaei, Hossein A Rahmani, and Yashar Deldjoo. 2022. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 770–779.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, Chichester, UK.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.

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*, pages 849–857.

Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip S Yu, and Peng Cui. 2022. Invariant preference learning for general debiasing in recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1969–1978.

Weifeng Zhang, Ting Zhong, Ce Li, Kunpeng Zhang, and Fan Zhou. 2022. Causalrd: A causal view of rumor detection via eliminating popularity and conformity biases. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1369–1378. IEEE.

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*, pages 3465–3476.

Ye Zhang and Byron Wallace. 2017. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proc. IJCNLP (Long Papers)*, pages 253–263, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ziyi Zhou, Xiaoming Zhang, Litian Zhang, Jiacheng Liu, Xi Zhang, and Chaozhuo Li. 2024. Fine-fake: A knowledge-enriched dataset for fine-grained multi-domain fake news detecction. *arXiv preprint arXiv:2404.01336*.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022a. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen

Zhuang. 2022b. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7178–7191.

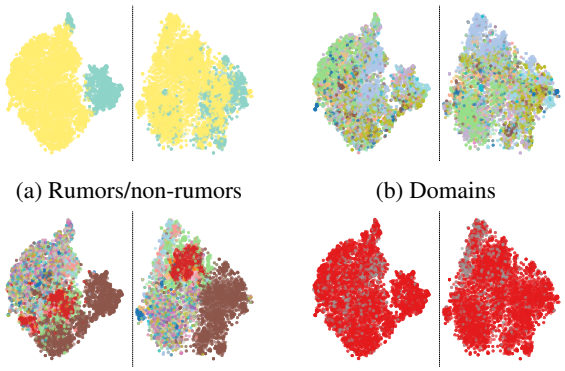
Yuchang Zhu, Jintang Li, Yatao Bian, Zibin Zheng, and Liang Chen. 2024. One fits all: Learning fair graph neural networks for various sensitive attributes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4688–4699.

A Appendix

A.1 Baselines and Implementations

Baselines The detailed baseline descriptions are as follows:

- BERT (Devlin et al., 2019): Utilizes transformer encoders to learn bidirectional text representations.
- EANN: Employs adversarial learning to remove event-specific information, enabling generalization to rumors from different events.
- MDFEND (Nan et al., 2021): Uses a mixture-of-experts architecture to explicitly leverage domain information for assisting multi-domain rumor detection.
- M3FEND (Zhu et al., 2022b): Utilizes a memory bank to store rumors with similar textual characteristics and domains, and incorporates three different views to learn robust textual representations of rumors.
- DualEmo (Zhang et al., 2021): Detects fake news by identifying the consistency of emotions between source news content and corresponding comments.
- dFEND (Shu et al., 2019): Leverages informative comments associated with a news content for rumor detection.
- ENDEF (Zhu et al., 2022a): Aims to remove entity bias observed in rumor detection training datasets to improve model robustness.
- CDD (Chen et al., 2023): Seeks to eliminate psycholinguistic bias in training data using predefined categories of emotional words.
- DTDBD (Li et al., 2024b): Utilizes a dual knowledge distillation architecture to learn informative representations of multi-domain rumors and applies adversarial training to mitigate domain bias.



(a) Rumors/non-rumors (b) Domains

(c) Platforms (d) Authors

Figure 10: Visualization of learned features on Chinese dataset. Left: MDFEND w/ our method. Right: MDFEND. Colors denote different groups.

- DeepSeek (Liu et al., 2024): The current state-of-the-art open-source LLM; we use the 7B version.
- GPT-4o (Achiam et al., 2023): The current state-of-the-art closed-source LLM.

Implementation Details Our method is implemented using the PyTorch framework. Unless otherwise specified in the main paper, we adopt 'bert-base-uncased' as the base model. To retain pre-trained knowledge, we freeze the first five layers of BERT. We use AdamW as the optimizer to update the weights of the entire framework. A pseudo algorithm of our framework is summarized in Algorithm 1, and hyperparameter settings not described in the main paper are provided in Table 5. For baseline implementations, we follow the released public code to replicate each framework and carefully tune the hyperparameters to achieve optimal performance. For fine-tuning LLMs, we apply the efficient parameter tuning technique LoRA for DeepSeek, while for GPT-4o, we utilize its official API service. All experiments are conducted on a single NVIDIA A100 GPU with 80GB of memory.

Hyperparameter	English	Chinese
Learning Rate	3.08×10^{-5}	8.67×10^{-6}
Weight Decay	3.94×10^{-6}	3.55×10^{-6}
Embedding Size	512	768
Batch Size	8	8
Early Stopping Patience	3	4

Table 5: Hyperparameter settings for English and Chinese datasets

A.2 Visualized Features on Chinese Dataset

The visualized features using the original MDFEND and our method on the Chinese dataset are

Algorithm 1 Training Algorithm

Require: Training data $\{(x_i, y_i)\}_{i=1}^{m}$, base model Φ, classifier f, hyperparameters k, λ

Initialize Φ, f; $\mathbb{A} \leftarrow \emptyset$
for each epoch **do**
  // Unfair partition discovery
  Freeze Φ; build partition $A \in \{0,1\}^{m \times k}$ via $\hat{s} = FC(\Phi(x))$
  Compute partition loss $\mathcal{L}_{par}$ (Eq. 2); store $(A, \mathcal{L}_{par})$
  Apply elbow method on $\{\mathcal{L}_{par}\}$ to retain optimal partitions $\mathbb{A}$
  // Invariant learning
  Unfreeze Φ; for each $A^i \in \mathcal{A}$ compute invariant loss $\mathcal{L}_{inv}^i$ (Eq. 3)
  Compute weights $\alpha_i = \exp(\mathcal{L}_{inv}^i) / \sum_j \exp(\mathcal{L}_{inv}^j)$
  Update Φ, f by minimizing $\sum_i \alpha_i \mathcal{L}_{inv}^i$
end for
return Φ, f

depicted in Fig. 10, where similar observations to those in the English dataset can be readily drawn.

A.3 Detailed OOD test results

We provide the detailed OOD test results in Table 6, from which it can be seen that our method is generally more robust and exhibits less performance variance across multiple experimental runs.

A.4 Dataset Statistics

We list the detailed dataset statistics in Table 7.

Setting	Lang.	MDFEND	DualEmo	CDD	GPT–4o	FIRM
Domain OOD (25% held out)	EN	74.20 ± 1.67	75.40 ± 2.31	74.80 ± 2.32	76.00 ± 1.12	77.20 ± 0.65
	ZH	81.40 ± 2.56	82.30 ± 1.79	81.40 ± 2.30	81.90 ± 1.11	84.60 ± 0.93
Domain OOD (50% held out)	EN	65.62 ± 1.26	69.13 ± 2.34	66.13 ± 2.12	68.65 ± 1.28	72.80 ± 0.73
	ZH	80.50 ± 2.43	79.60 ± 2.14	80.00 ± 1.68	80.80 ± 1.24	82.50 ± 0.84
Platform OOD (25% held out)	EN	68.71 ± 20.66	69.54 ± 15.40	65.34 ± 16.80	71.30 ± 14.50	75.40 ± 10.20
	ZH	79.20 ± 9.80	76.40 ± 10.30	78.30 ± 11.60	75.40 ± 9.70	81.40 ± 6.80
Platform OOD (50% held out)	EN	67.65 ± 18.54	64.34 ± 19.52	65.53 ± 18.80	65.55 ± 16.40	71.23 ± 12.30
	ZH	69.06 ± 1.42	71.34 ± 2.34	73.12 ± 2.56	72.49 ± 1.21	77.63 ± 0.86

Table 6: Out-of-distribution (OOD) accuracy (%) on English (EN) and Chinese (ZH) datasets. Each entry shows mean $\pm$ standard deviation.

Dataset	Chinese	English
Rumors (Fake News)	17,895	9,407
Non-rumors (True News)	6,074	7,502
Total	23,969	16,909
Rumor Ratio by Topic		
Health	0.850	0.428
International/Conflict	0.962	0.497
Lifestyle/Society	0.883	0.637
Politics	0.870	0.519
Technology	0.660	-
Environment/Energy	0.864	-
Disasters/Accidents	0.817	-
Entertainment/Sports	0.561	0.589
Social Life	0.642	-
Military	0.867	-
Education/Exams	0.579	-
Finance/Business	0.437	0.521
Uncategorized	-	0.602
Rumor Ratio by Platform		
Taiwan FactCheck Center	0.928	-
Weibo	0.492	-
China Internet Joint Rumor Refutation Platform	0.977	-
Tencent News Fact-check Platform	0.926	-
China Daily	1.000	-
Mygopen	0.933	-
HKBU Fact Check	0.961	-
HKU Annie Lab	0.951	-
AFP Fact Check	1.000	-
Factcheck Lab	0.903	-
NetEase	0.846	-
Phoenix	0.849	-
Tencent	0.865	-
Sohu	0.843	-
Baijiahao	0.856	-
Sina	0.862	-
Interface News	0.875	-
CDC.gov	-	0.011
AP News	-	0.093
CNN	-	0.103
Reddit	-	0.763
Twitter	-	0.621
Washington Post	-	0.323
Snopes	-	0.673
Rumor Ratio by Author		
Unknown	0.730	0.632
Known	0.872	0.486

Table 7: Dataset Statistics for Chinese and English News