Beyond Content: Integrating Generated User Intent and Planned Behavior Theory for Robust Fake News Detection

Anonymous ACL submission

Abstract

The dissemination of true and fake news is often driven by distinct user motivations, yet existing detection methods predominantly focus on news content or propagation structures, often overlooking these underlying intents. This oversight can make such methods vulnerable to sophisticated adversarial strategies, such as crafted fake content or deceptive user engagement. While large language models (LLMs) provide rich, multi-dimensional behavioral insights, their standalone performance in detection often lags behind supervised models. To bridge this gap, we propose a novel computational framework that integrates the Theory of Planned Behavior (TPB) with LLM-generated user intent, enabling a deeper understanding of users' decision-making processes in news sharing. We employ a two-layer contrastive feature fusion mechanism to construct comprehensive behavioral representations, significantly enhancing fake news detection. Extensive experiments across four diverse datasets demonstrate that our method also exhibits remarkable robustness against adversarial attacks.

1 Introduction

In the era of large language models (LLMs), information fabrication is becoming increasingly sophisticated (Huang and Sun, 2024; Lucas et al., 2023). Traditional detection approaches have primarily relied on semantic or stylistic features extracted from news content (Vlachos and Riedel, 2014; Wu et al., 2020a). However, unlike traditional fake news, which often displays discernible inconsistencies, LLM-generated content exhibits human-like coherence and adaptability (Sun et al., 2024), which undermines the performance of detectors (Sadasivan et al., 2024). Experiments have shown that data rewritten by LLMs can lead to a reduction of up to 38.3% in the F1 score (Wu et al., 2024).

Therefore, recent fact-checking research has attempted to identify deceptive evidence from related sources, such as user comments (Shu et al., 2019) and relevant articles (Wu et al., 2020b). Social engagement-based approaches then introduce propagation features such as diffusion graphs (Bian et al., 2020) and propagation patterns (Sun et al., 2023). While these strategies have enhanced the robustness of detection, they remain susceptible to malicious social manipulation (Wang et al., 2023), such as publishing extremist comments. Studies have shown that social attacks can achieve a success rate as high as 90% , particularly against models that rely on networks (Wang et al., 2023), highlighting the pressing need for more robust frameworks that can identify manipulative clues.

The key to identifying manipulative behavior lies in finding anomalous actions, a task that has become increasingly challenging. Given the differing intentions behind real and fake news (Wang et al., 2024b), users' intention for sharing news also varies, with real news often shared to express opinions or convey information, while the sharing of fake news is more likely influenced by emotional factors (McLoughlin et al., 2024). However, since users' decision-making processes are always complex and unobservable, it is particularly difficult for traditional deep learning models to understand user intent. Therefore, LLMs with their reasoning capabilities, have been widely applied in detection tasks. Detection achieved through standardized prompt learning is not always satisfactory (Wang et al., 2024a), advanced strategies such as workflow optimization and retrieval-augmentation have been shown to improve performance (Li et al., 2024a; Cheung and Lam, 2023). Studies utilizing generative comments (Nan et al., 2024) or simulated social engagement (Wan et al., 2024) to augment data have produced richer refutation evidence, but also introducing non-real data to the system, which may increase the complexity of data processing.

Therefore, in the increasingly chaotic landscape of social media, uncovering the underlying inten-

tions behind user behavior and constructing accurate behavioral models for reliable and robust fake news detection representcritical challenges. In this paper, we leverage the reasoning capabilities of LLMs to infer users' underlying intentions on social media, and develop a computational framework based on the Theory of Planned Behavior (TPB) (Ajzen, 1991), a framework from social psychology that explains how attitudes, subjective norms, and perceived behavioral control shape users' intentions and behavior, to describe users' behavioral planning process on online social media. By combining contrastive learning with feature fusion techniques, we achieve rich representations of user behavior for robust fake detection.

Specifically, as illustrated in Fig.1, users' behavioral planning process is always unobservable; therefore, based on the observed social media environment: news content, user attributes, and user behavior, we first categorize users' attributes into three dimensions: basic profiles, social traits, and historical posts. These features, along with user behaviors and news content, are processed by LLM to obtain the inferred users' intention. Then, we employ computational methods to map the initial variables of user attitudes, subjective norms, and perceived behavioral control from the non-contact social data. Using these three variables, we predict users' intention and perform the first layer of fusion. To bridge the inferred intentions from the LLM with the TPB-based predictions, we employ a contrastive loss, ensuring alignment between the two perspectives. Subsequently, user behaviors are predicted based on the intentions and perceived control, with predictions iteratively compared to observed real behaviors to refine the process. Finally, the enhanced fusion of intentions and behaviors is utilized as the ultimate representation for detection, thereby enhancing performance and improving resilience against adversarial attacks. Our contributions can be summarized as:

- Innovative Intention Inference with LLMs : We employ LLMs to uncover the underlying intentions driving user action, thus facilitating a deeper understanding of anomalies and revealing the motivational differences between true and fake news.

- Interpretable Behavior Modeling via TPB: By developing the computational TPB for online information diffusion, we bridge the gap between psychological theory and computational frameworks, enhancing both the interpretability of feature fusion and its robustness in complex decision-making scenarios.

- Robust Detection in Adverse Scenarios: Experiments on 4 datasets reveal that combining LLM-inferred intent with TPB-guided modeling significantly improves detection accuracy and maintains stable in extreme scenarios of data scarcity and adversarial attacks.

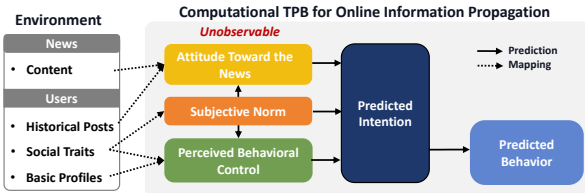


Figure 1: Computational Theory of Planned Behavior (TPB) for information spreading behavior modeling.

2 Related Work

2.1 Social context-enhanced Fake Detection

Detection methods relying solely on news content are increasingly inadequate for high-quality fake news detection. Consequently, recent research has incorporated relevant social context and propagation features into detection frameworks. For instance, dEFEND(Shu et al., 2019) constructs a news-comment network, leveraging the semantic correlations between content and comments for detection. The DualEmo model (Zhang et al., 2021) analyzes the emotional characteristics and the emotional gap between news content and user responses to enhance detection. Bian et al.(Bian et al., 2020) designed a Bi-GCN model to capture the bidirectional propagation patterns. GCNFN (Monti et al., 2019) leverages users' profiles to supplement comment embeddings, while UPFD (Dou et al., 2021) further incorporates users' historical posts to capture their intrinsic preferences. HG-SL (Sun et al., 2023) constructs a hypergraph based on users' sharing behavior and incorporates statistical propagation features to enhance the learning. Similarly, HGFND(Jeong et al., 2022) constructs hypergraphs from different perspectives, effectively capturing the dissemination patterns of news.

Instead of using real comments, GenFEND (Nan et al., 2024) utilizes LLMs as a user simulator and comment generator, generates comments from potential users with diverse profiles. DELL (Wan

et al., 2024) simulates the whole social system with different user engagements and introduces 6 proxy tasks to enhance the news understanding. These strategies enrich social data but also introducing non-real data to the system, which may increases the complexity of data mining.

2.2 User Behavior Modeling in Information Propagation

User motivations for information dissemination are influenced by multiple factors, including news attributes (e.g., sentiment) (Horner et al., 2021), with fake news often eliciting stronger emotional reactions such as anger (McLoughlin et al., 2024). Moreover, users attributes and their social circles also influence their behavior. Altay et al. (Altay et al., 2022) found that users with more friends share less fake news, Cheng et al.(Cheng et al., 2021) modeled the unbiased fake news propagation and revealed that verified users or users have more tweets are less likely to be suspicious. Gimpel et al. (Gimpel et al., 2021) found that fake news is frequently shared within trusted social circles.

Since user's behavioral decision-making process is unobservable, theoretical guidance is essential (Zhang et al., 2022). Social identity theory reveals that users tend to conform to the viewpoints prevalent within their community to achieve a sense of belonging (Dmj et al., 2018). The Theory of Planned Behavior (TPB) (Ajzen, 1991) explains and predicts human behavior based on three key components: attitudes, subjective norms, and perceived behavioral control. Attitudes refer to an individual's evaluation of a behavior, subjective norms represent the perceived social pressure and perceived behavioral control reflects the individual's confidence to execute the behavior. Together, these factors shape intentions, which in turn lead behavior. TPB is widely used across disciplines, it also helps reveal how personal evaluation (attitude) and social influence (subjective norm) intersect with an individual's sense of agency (perceived control) influence user's sharing behavior.

3 The Proposed Method

We design a computational framework based on the Theory of Planned Behavior (**TPB**) with LLM-generated user **Intent** (**TPB-Intent**), as shown in Fig.2, which consists of 3 major components: Data Filtering and Processing, LLM-based User Intention Inference, and TPB-based Computational User Behavior Learning and Fake Detection.

3.1 Problem Formulation

Given a list of m news $N = \{n_1, n_2, ..., n_m\}$, for each news $n_i = (d_i, g_i)$, we have the text of the news content d_i and user behavior records $g_i = ((u_1, b_{i,1}), (u_2, b_{i,2})), ..., (u_n, b_{i,n}))$, where u_j is a user participating in n_i's propagation, and $b_{i,j}$ represents the behavior of u_j in the propagation. For each user, we divide the user's personal information into basic profiles pro_j, social traits soc_j and historical posts his_j, thus $u_j = (pro_j, soc_j, his_j)$. Each news is assigned a label $y_i \in \{0, 1\}$, if news n_i is fake, $y_i = 1$, otherwise $y_i = 0$. Our model aims to find the intent $int_{i,j}$ of user u_j in n_i, and predict a label $\hat{y}_i$ for n_i.

3.2 Data Filtering and Processing

3.2.1 Top-K Influential Nodes Identification

Considering the high computational cost of utilizing LLMs and the observation that key users often play a pivotal role in shaping overall propagation, we first identify influential user nodes from the propagation network. We calculate 7 network metrics (Node depth, Children count, Total reach, Response latency, Propagation duration, Degree centrality and Betweenness centrality) for each node, and sort them to select the k nodes with the highest ranking as high-influence nodes.

3.2.2 User data processing

The raw user data, comprising both numerical and textual information, is challenging for LLMs to interpret directly. To address this, we categorize numerical data into distinct levels based on statistical distributions. For example, the number of user posts is segmented into five levels (e.g., few, relatively few, moderate, relatively many, and many) and converted into descriptive statements like "The user has a (level) volume of posts" (see Fig.2). This approach enhances embedding learning by introducing semantic granularity and improves LLMs' interpretability by transforming quantitative data into meaningful linguistic representations.

3.3 LLM-based User Intention Inference

Capturing the diverse, hidden, and complex nature of user intentions is challenging for traditional deep learning methods. Leveraging the knowledge and reasoning capabilities of large language models (LLMs), we propose a novel approach to reconstruct user intentions by integrating user attributes,

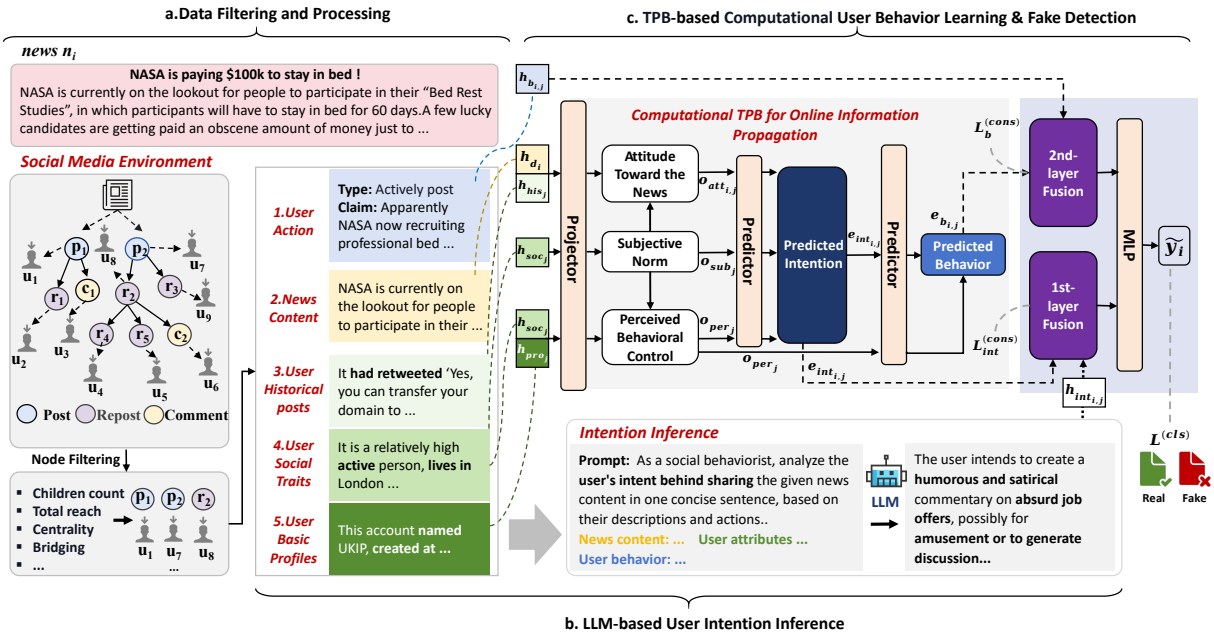


Figure 2: The overall architecture of TPB-Intent: a.Data filtering and processing module for influential nodes identification and user data processing, b.LLM-based user intention inference module to infer users' intentions behind behavior, and c.TPB-based computational user behavior learning module for feature fusion and detection.

news content, and behavioral data through strategic prompting. This enables the discovery of intention differences in the spread of true and false information and addresses data gaps in the intention component of the Theory of Planned Behavior (TPB). Guided by TPB, which posits that user actions are driven by intentions, we use LLMs to infer intentions from actual behaviors, news content, and user characteristics. To handle the complexity of intentions, we employ an open-ended prompting strategy, generating concise, single-sentence interpretations of user intent without predefined categorical constraints. For each news n_i, we leverage LLM to infer the intentions of the identified k key users. The inferred intention is denoted as $int_{i,j}$ for u_j in the spread of n_i.

Intention Inference Prompt

System Prompt: *As a social behaviorist, analyze the users intent behind sharing the given news content in one concise sentence based on their descriptions and actions,using the Theory of Planned Behavior.*
Note: Focus solely on the user's intent,considering the authenticity of the news, and account for the distinct motivations when sharing true versus fake news.
Context Input:
News content: [d_i]
User behavior type and response: [b_j]
User description: [pro_j], [soc_j]

3.4 TPB-based Computational User Behavior Learning and Fake Detection

Due to variable omissions, the original TPB cannot be directly applied to online information dissemination scenarios. Therefore, we propose a three-step approach to model this process: (1) map observed social data to the initial variables of TPB; (2) perform computational intention learning and contrastive fusion for enhanced intentions; and (3) conduct computational behavior learning and contrastive fusion for enhanced behaviors.

3.4.1 Mapping Social Data to TPB Constructs

For all raw data, we use BERT(Devlin et al., 2019) as the encoder. We first use the encoder to obtain the representation of user historical posts h_{his_j}, social traits h_{soc_j}, profiles h_{pro_j} and news content h_{d_i}, each $h \in R^{d'}$ is a d'-dimensional vector. Then, we employ a feature projector $\mathcal{P}(\cdot)$ to map the raw features into the elements of the TPB framework. This module uses a two-layer fully connected network and combines activation functions and regularization operations. Since a user's attitude toward participating in the spread of a news topic is influenced by their prior cognition and the content of the news, we combine the text embeddings of the news with the user's historical posts and feed them into the projection layer to calculate the attitude. Similarly, subjective norms mainly reflect the influence of others on users, so we use the user's social attributes to measure this factor. Perceived behavioral control represents the user's overall sense of self-efficacy in forming and achieving their intentions. This aspect can be inferred from their self-cognition and

social preference. Thus, we integrated the user's profiles and social traits to compute their potential perceived behavioral control ability. Therefore, we obtain the projected embeddings for attitudes $\mathbf{o}_{att_{i,j}} \in R^d$, subjective norms $\mathbf{o}_{sub_j} \in R^d$ and perceived behavioral control $\mathbf{o}_{per_j} \in R^d$:

$$\mathbf{o}_{att_{i,j}} = \mathcal{P}(\mathbf{h}_{his_j} + \mathbf{h}_{d_i}), \quad (1)$$

$$\mathbf{o}_{sub_j} = \mathcal{P}(\mathbf{h}_{soc_j}), \quad (2)$$

$$\mathbf{o}_{per_j} = \mathcal{P}(\mathbf{h}_{pro_j} + \mathbf{h}_{soc_j}). \quad (3)$$

3.4.2 Computational Intent Learning and Contrastive Fusion

Subsequently, we concatenate $\mathbf{o}_{att_{i,j}}$, $\mathbf{o}_{sub_j}$ and $\mathbf{o}_{per_j}$, use the combined representation to predict the intent of the user u_j in spreading news n_i. Given the complexity of intent formation, which involves deep logical and semantic transformations, the objective of this step is not to directly predict the intent. Instead, we aim to approximate it in a comparative manner, bringing the representation relatively closer to the LLM inferred intent $\mathbf{h}_{int_{i,j}}$. To achieve this, we first input the combined embeddings into a linear layer for feature refinement, yielding an intermediate variable $\mathbf{e}_{int_{i,j}}$:

$$\mathbf{e}_{int_{i,j}} = \mathbf{W_1}[\mathbf{o}_{att_{i,j}}, \mathbf{o}_{sub_j}, \mathbf{o}_{per_j}] + \mathbf{b}_1 \quad (4)$$

where $\mathbf{W_1}$ is weight matrix, $\mathbf{b}_1$ is bias vector.

Then, We compute a contrastive loss to optimize the embedding $\mathbf{e}_{int_{i,j}}$, ensuring it is relatively close to the embedding of LLM inferred intention $\mathbf{h}_{int_{i,j}}$ of the same user-news pair, while remaining relatively distant from other pairs:

$$\mathcal{L}_{int}^{cons}(\mathbf{e}_{int_{i,j}}, \mathbf{h}_{int_{i,j}}) = -\frac{\exp\left(\cos\left(\mathbf{e}_{int_{i,j}}, \mathbf{h}_{int_{i,j}}\right)/\tau\right)}{\sum_k \exp\left(\cos\left(\mathbf{e}_{int_{i,j}}, \mathbf{h}_{int_k}\right)/\tau\right)}. \quad (5)$$

where $k = (p, q)$ is other candidate user-news pairs,τ is the temperature scaling parameter, $cos(\cdot)$ represents the cosine similarity.

Finally, we integrate the predicted $\mathbf{e}_{int_{i,j}}$ and the inferred $\mathbf{h}_{int_{i,j}}$ to obtain an enhanced intent representation for the first layer of feature fusion:

$$\mathbf{z}_{int_{i,j}} = \mathbf{W_2}[\mathbf{e}_{int_{i,j}}, \mathbf{h}_{int_{i,j}}] + \mathbf{b}_2 \quad (6)$$

3.4.3 Computational Behavior Learning and Contrastive Fusion

Based on the original TPB, an individual's perceived behavioral control, together with their intention, drives the execution of specific behaviors. In this process, the behavioral control plays a pivotal role in motivating users to take concrete actions rather than merely having intentions. If an individual's behavioral control is weak, they may perceive that their actions will not achieve the intended outcome, leading to a potential abandonment of action.

In this step, we begin by projecting the user's real action on social media to obtain behavior embeddings $\mathbf{h}_{b_{i,j}}$. Subsequently, we concatenate and transform the predicted user intent $\mathbf{e}_{int_{i,j}}$ with the user's perceived behavioral control embeddings $\mathbf{o}_{per_j}$ to obtain the predicted behavior $\mathbf{e}_{b_{i,j}}$. Finally, the real behavior embeddings and the predicted embeddings are integrated to derive $\mathbf{z}_{b_{i,j}} \in R^d$:

$$\mathbf{e}_{b_{i,j}} = \mathbf{W}_3[\mathbf{e}_{int_{i,j}}, \mathbf{o}_{per_j}] + \mathbf{b}_3, \quad (7)$$

$$\mathbf{z}_{b_{i,j}} = \mathbf{W}_4[\mathbf{e}_{b_{i,j}}, \mathbf{h}_{b_{i,j}}] + \mathbf{b}_4. \quad (8)$$

To optimize behavioral learning, similarly, we compute a contrastive loss to encourage $\mathbf{e}_{b_{i,j}}$ to be close to $\mathbf{h}_{b_{i,j}}$ and away from $\mathbf{h}_{b_k}$ of other pairs.

$$\mathcal{L}_b^{cons}(\mathbf{e}_{b_{i,j}}, \mathbf{h}_{b_{i,j}}) = -\frac{\exp\left(\cos\left(\mathbf{e}_{b_{i,j}}, \mathbf{h}_{b_{i,j}}\right)/\tau\right)}{\sum_k \exp\left(\cos\left(\mathbf{e}_{b_{i,j}}, \mathbf{h}_{b_k}\right)/\tau\right)}. \quad (9)$$

3.4.4 Fake Detection

Ultimately, the representations $\mathbf{z}_{int_{i,j}}$ and $\mathbf{z}_{b_{i,j}}$, derived from the two levels of learning, are fed into a Multi-Layer Perceptron (MLP) to compute the suspicion score for news articles.

$$\hat{y}_i = \mathrm{MLP}(\mathbf{Z}_{int_i}) + \mathrm{MLP}(\mathbf{Z}_{b_i}) \quad (10)$$

The final loss function incorporates both the contrastive loss for mapping intents and behaviors during the learning process and the classification loss.

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right) \quad (11)$$

$$\mathcal{L} = \lambda\mathcal{L}_{cls} + (1 - \lambda)\left(\mathcal{L}_{int}^{cons} + \mathcal{L}_b^{cons}\right) \quad (12)$$

λ determines the contribution of losses.

4 Experiments

We conduct extensive experiments on four widely used datasets to answer the questions: (**RQ1**)Does TPB-Intent outperform baselines in fake detection? (**RQ2**)Are intentions identified by LLMs more valuable than real data? (**RQ3**) What are the most important strategies and features in the model?(**RQ4**)How robust is TPB-Intent to data constraints and adversarial attacks?

Table 1: Performance comparison on all datasets, where the best and second-best results are highlighted in bold and underlined, respectively. A:accuracy, P:precision, R:recall.

Method	Politifact				Gossipcop				Mcfend				Weibo21			
	A	P	R	macF1	A	P	R	macF1	A	P	R	macF1	A	P	R	macF1
News content-based																
EANN	0.728	0.709	0.716	0.714	0.694	0.701	0.687	0.692	0.638	0.626	0.543	0.540	0.873	0.874	0.873	0.874
SentGCN	0.848	0.857	0.840	0.840	0.728	0.703	0.682	0.698	0.593	0.583	0.523	0.528	0.878	0.878	0.878	0.879
BERT	0.861	0.861	0.847	0.852	0.770	0.749	0.710	0.721	0.708	0.692	0.625	0.625	0.874	0.874	0.874	0.874
RoBERTa	0.850	0.851	0.838	0.842	0.779	0.773	0.707	0.722	0.690	0.702	0.597	0.578	0.894	0.894	0.895	0.894
Social context-enhanced																
dEFEND	0.899	0.901	0.898	0.895	0.914	0.907	0.896	0.901	0.766	0.749	0.737	0.739	0.924	0.924	0.924	0.924
DualEmo	0.928	0.931	0.925	0.924	0.918	0.917	0.911	0.913	0.762	0.741	0.728	0.725	0.940	0.940	0.940	0.940
BiGCN	0.907	0.906	0.902	0.903	0.919	0.913	0.906	0.918	0.750	0.733	0.702	0.706	0.932	0.933	0.932	0.932
UPFD-prefer	0.819	0.813	0.815	0.813	0.936	0.926	0.932	0.928	-	-	-	-	-	-	-	-
UPFD-profile	0.855	0.854	0.842	0.846	0.935	0.924	0.931	0.927	0.738	0.729	0.671	0.670	0.748	0.839	0.772	0.699
LLM-enhanced																
ChatGPT-n	0.731	0.938	0.581	0.717	0.757	0.674	0.517	0.585	0.446	0.605	0.411	0.489	0.787	0.846	0.694	0.762
Claude-n	0.721	**0.964**	0.545	0.697	0.767	0.816	0. 384	0.5220	0.418	0.630	0.238	0. 346	0.662	0.767	0.448	0.565
L-Defense	0.919	0.915	0.916	0.915	0.839	0.813	0.816	0.814	0.701	0.642	0.642	0.642	0.911	0.912	0.912	0.911
GenFEND	0.908	0.907	0.906	0.907	0.922	0.928	0.906	0.914	0.766	0.748	0.732	0.728	0.918	0.920	0.918	0.918
Ours	**0.955**	0.954	**0.954**	**0.953**	**0.952**	**0.945**	**0.947**	**0.946**	**0.800**	**0.784**	**0.774**	**0.771**	**0.950**	**0.950**	**0.949**	**0.950**

4.1 Experimental Setup

4.1.1 Datasets

We conduct experiments on four datasets, including two English datasets (Politifact and Gossipcop) and two Chinese datasets (Mcfend and Weibo21). Politifact focuses on political news, while GossipCop primarily covers entertainment news. Mcfend aggregates data from multiple fact-checking websites, Weibo21 is a multi-domain dataset from Weibo. The detailed statistics are provided in Appendix A

4.1.2 Baselines

We compare our model with 12 baselines in three categories(check Appendix B for details): News content-based methods (EANN, SentGCN, BERT, RoBERTa) rely solely on the textual content of news. Social context-enhanced approaches (dEFEND, DualEmo, BiGCN, UPFD) incorporate user engagement data. LLM-enhanced methods (ChatGPT, Claude, L-Defense, GenFEND) leverage inference and generation capabilities of LLMs.

4.1.3 Evaluation Metrics and Settings

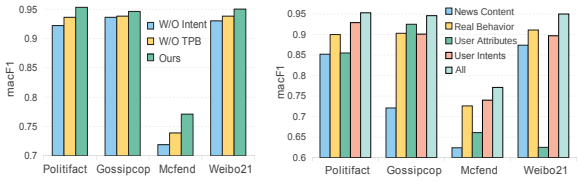
We use the accuracy(**A**), precision(**P**), recall(**R**) and **F1** score for evaluation. Our experiments are conducted on a 12 GB GeForce GTX 2080Ti GPU. We use 5-fold cross validation to evaluate models' performance. For Politifact and Mcfend, the number of high-influence users retained (k) is set to 50. For Gossipcop and Weibo21 with more news, k is set to 20. As Mcfend and Weibo21 lack user historical posts, we replace h_{his_j} with profile and social traits ($h_{pro_j} + h_{soc_j}$) in Eq.1. For baselines, we retain their settings. For our model, we prompt ChatGPT (gpt-4o-mini) to infer users' intent, implement it in PyTorch and adopt Adam as the optimizer, train 50 epochs to obtain best performance. The learning rate is 0.001 and the batch size is 32. The dimension of learned representations d is 64.

4.2 Results on Fake News Detection (RQ1)

Table1 presents model performance across four datasets. Key observations include: (1) Our model achieves the highest overall performance, demonstrating the effectiveness of integrating LLM-generated user intent with the theory of planned behavior. DualEmo performs competitively on Politifact and Weibo21 (macF1: 0.924, 0.940), emphasizing the role of emotional features. On Gossipcop, user features are critical, as UPFD-prefer (leveraging user posting history) reaches a macF1 of 0.928. For Mcfend, semantic information in user comments is decisive, with dEFEND and GenFEND outperforming other baselines. (2) Among content-based methods, RoBERTa and BERT consistently achieve higher scores, reflecting their strong semantic modeling capabilities. However, their reliance on news content alone limits performance, particularly on datasets where social context is crucial (Mcfend, Gossipcop). (3) Social context-enhanced methods generally outperform content-based approaches, highlighting the value of user interaction features. (4) LLM-enhanced models show mixed results. Standalone LLMs perform

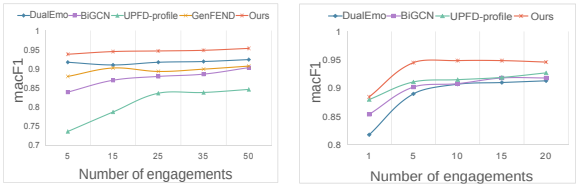
Table 2: Performance comparison after replacing user comments or attributes with inferred intent.

Method	Politifact				Gossipcop				Mcfend				Weibo21			
	A	↑ %	**macF1**	↑ %	**A**	↑ %	**macF1**	↑ %	**A**	↑ %	**macF1**	↑ %	**A**	↑ %	**macF1**	↑ %
dEFEND	0.912	1.45	0.915	2.23	0.918	0.44	0.911	**1.11**	0.779	1.7	0.747	1.08	0.915	-0.98	0.915	-0.98
DualEmo	0.933	0.5	0.930	0.65	0.922	0.44	0.918	0.55	0.779	2.23	0.745	2.76	0.928	-1.28	0.928	-1.28
BiGCN	0.927	2.18	0.926	2.55	0.930	**1.2**	0.927	0.98	0.762	1.6	0.725	2.69	0.921	-1.18	0.921	-1.18
UPFD-Pro	0.891	**4.21**	0.889	**5.08**	0.920	-1.71	0.916	-1.29	0.763	**3.39**	0.735	**9.7**	0.921	**23.13**	0.921	**31.76**

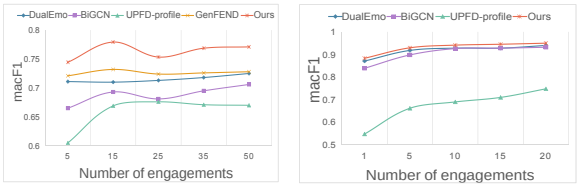


(a) Ablation of strategies (b) Ablation of features

Figure 3: Results of strategy and feature ablation.



(a) Politifact (b) Gossipcop



(c) Mcfend (d) Weibo21

Figure 4: Impact of engagement data volume.

poorly across most datasets, particularly on Mcfend (macF1: 0.411, 0.346), underscoring the necessity of task-specific adaptations.

4.3 Impact of Intent on Fake Detection (RQ2)

To validate that inferred user intentions may provide richer cues than real data, we adapted existing social context-enhanced models by substituting user comments or attributes with inferred intentions. Results in Table 2 show that while replacing comments on Weibo21 and user attributes on Gossipcop led to minor declines, inferred intent generally enhanced model performance, especially in cases with incomplete user features (e.g., macF1 increased by 31.76% for Weibo21 under UPDF-Pro(file)). This underscores the potential of intent as a valuable supplementary cue for detection.

4.4 Ablation Study (RQ3)

As shown in Fig.3(a), removing inferred intent causes a significant performance drop across all datasets, particularly on Mcfend (macF1: 0.77 → 0.72), highlighting the critical role of inferred intent in enhancing detection. Similarly, excluding TPB-guided aggregation and instead summing individual features at the same level results in suboptimal macF1 scores, demonstrating the effectiveness and theoretical grounding of TPB in feature fusion.

The feature ablation (Fig.3(b)) shows that our approach effectively integrates all features to achieve highest macF1. In Politifact and Mcfend, user intentions play a dominant role, aligning with the nature of political or social events, where the drivers of true and fake news dissemination differ significantly. For Weibo21, users' real behavior emerges as the most critical feature, likely indicating that users' response always contain direct evidence. For Gossipcop, user attributes prove to be significant, indicating less distinct intentions in entertainment domain, making account information essential.

4.5 Robustness Analysis (RQ4)

4.5.1 Data Limitation

Fig.4 shows the performance of social-context enhanced models under varying user engagement constraints. Our model maintains robust performance across all datasets. This stems from its integration of intent and the comprehensive feature fusion, enabling effective representation learning even with minimal interactions. While competing models also improve with more engagements, UPFD-profile struggles under sparse interactions, likely due to the limited semantic richness of user profiles. BiGCN, heavily reliant on networks, also suffers from restricted interactions.

4.5.2 News Style Attack

For news style attacks, we follow (Wu et al., 2024), modifying the stylistic presentation of both fake and real news. For the English datasets Politifact, we instruct the LLM (chatgpt-4o-mini) to rewrite

Table 3: Impact of news style attacks.

Method	Politifact				Mcfend			
	A	↓ %	macF1	↓ %	A	↓ %	macF1	↓ %
BERT	0.664	22.9	0.613	28.1	0.521	26.4	0.366	41.4
DualEmo	0.819	11.6	0.815	11.8	0.722	5.3	0.634	12.6
BiGCN	0.771	15.0	0.763	15.5	0.740	1.3	0.689	2.41
ChatGPT	0.460	37.1	0.441	38.5	0.404	9.4	0.370	24.3
Ours	**0.942**	**1.36**	**0.940**	**1.36**	**0.789**	**1.38**	**0.764**	**0.91**

Table 4: Impact of user engagement attacks.

Method	Politifact				Mcfend			
	A	↓ %	macF1	↓ %	A	↓ %	macF1	↓ %
DualEmo	0.921	0.75	0.917	7.57	0.787	-3.28	0.749	-3.25
BiGCN	0.931	**-2.65**	0.928	**-2.77**	0.851	-13.5	0.822	-16.4
UPFD-Pro	0.797	6.78	0.787	6.97	**0.954**	**-29.3**	**0.951**	**-41.9**
GenFEND	0.910	-0.22	0.908	-0.11	0.807	-5.35	0.779	-7.01
Ours	**0.948**	0.73	**0.947**	0.63	0.828	-3.50	0.796	-3.24

fake news in the style of "New York Times" and real news in the style of "The National Enquirer". For the Chinese dataset Mcfend, we direct LLM to rewrite fake news in the style of "People's Daily" in China, while real news is rewritten in an exaggerated, attention-grabbing style to mimic misleading content due to the lack of a suitable fake news proxy. The models are trained on the original data while the test set is replaced to simulate attacks.

Table 3 reports the performance degradation (↓ %) of various models under news style attacks. Our model only leverages news content when computing users' attitudes while incorporating multiple dimensions of features to supplement learning process, thus achieving the highest robustness. In contrast, content-only models (e.g., BERT, ChatGPT-n) suffer severe degradation (up to 41.44%), revealing their vulnerability to stylistic variations. DualEmo and BiGCN, which incorporate user engagement, emotions and network structures, exhibit moderate resistance, with degradation ranging from 11%–15% (Politifact) and 2%–12% (Mcfend).

4.5.3 User Engagements Attack

For user engagement attacks, we adopt a simplified version of the strategy from (Wang et al., 2023). We identify users engaging exclusively with real or fake news and simulate cross-interference using LLM (chatgpt-4o-mini). For real-only users, we generate their responses with positive stance and inferred intents for fake news; for fake-only users, we generate interactions with real news. To enhance realism, 5 interactions per news item are generated and randomly integrated into propagation structure.

Results in Table 4 show that most models experience anomalous accuracy increases under attacks, particularly on Mcfend dataset. This is likely because many fake news-sharing accounts are fact-checking accounts, whose interactions with true news provide positive cues. Moreover, the limited number of users consistently sharing one type of news leads to their repeated use in attacks, allowing models to detect patterns rather than being misled. This effect is most notable in UPFD-Pro(file), which relies heavily on user profiles, with its F1 fluctuating by 41.94%. In contrast, our model maintains stability by integrating intent and behavioral insights, reducing vulnerability to manipulation.

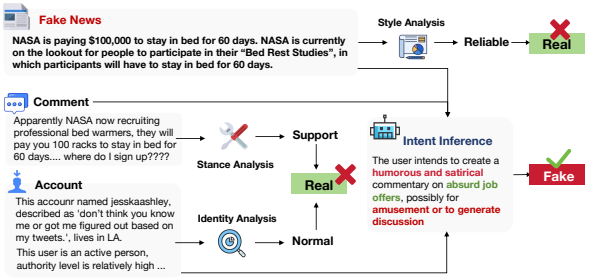


Figure 5: Case: intent reveals subtle deception.

4.6 Case Study

The case in Fig.5 highlights LLMs' ability to identify hidden intentions, such as sarcasm, in seemingly credible news posts and user responses. For instance, the fabricated post about NASA appears reliable due to its professional language, potentially misleading basic detectors into classifying it as "Real." Similarly, the user's excited comment might be misinterpreted as supportive by stance analyzers, leading to incorrect judgments. Our work shows that LLMs can uncover humor and sarcasm by analyzing social contexts, revealing intentions like "for amusement" or "generate discussion" rather than genuine support. This enables the detector to reassess the post's authenticity.

5 Conclusion

Our proposed framework integrates LLM-inferred user intent with the Theory of Planned Behavior to enhance fake news detection across multiple datasets, which not only improves detection performance but also strengthens robustness and interpretability, maintaining stable results even under various adversarial attack scenarios. This highlights the value of incorporating psychological and behavioral insights into computational models for fake news detection and mitigation.

6 Limitations

Despite the effectiveness of our approach, several limitations remain. First, while our model integrates user intent inference with the Theory of Planned Behavior (TPB), its performance will be affected by the accuracy of Large Language Models (LLMs) in capturing user intentions. Errors in intent inference could propagate through the framework, potentially impacting detection robustness. Second, our approach relies on social engagement features, which may be sparse or unavailable for certain news articles, limiting applicability in low-resource settings. Additionally, the model assumes that user behaviors align with their inferred intentions, which may not always hold due to strategic misinformation campaigns or adversarial manipulations. Future work will explore adaptive strategies to mitigate these issues, such as dynamic intent recalibration and cross-platform behavior modeling.

7 Ethical Consideration

We utilize publicly available datasets curated by previous researchers, strictly adhering to all relevant legal and ethical standards during data acquisition and usage. To mitigate potential societal risks, we provide only prompt templates without disclosing the specific content of LLM-generated intentions. This approach ensures responsible use of the technology while maintaining transparency in our methodology.

References

Icek Ajzen. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211. Theories of Cognitive Self-Regulation.

Sacha Altay, Anne-Sophie Hacquin, and Hugo Mercier. 2022. Why do so few people share fake news? it hurts their reputation. *New Media Soc.*, 24(6):1303–1324.

Anthropic. 2024. Claude. https://www.anthropic.com. Accessed: 2024-12-12.

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

Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 148–157, New York, NY, USA. Association for Computing Machinery.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023, Taipei, Taiwan, October 31 - Nov. 3, 2023*, pages 846–853. IEEE.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, pages 4171–4186. ACL.

Lazer Dmj, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, and Rothschild D. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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

Henner Gimpel, Sebastian Heger, Christian Olenberger, and Lena Utz. 2021. The effectiveness of social norms in fighting fake news on social media. *J. Manag. Inf. Syst.*, 38(1):196–221.

Christy Galletta Horner, Dennis F. Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2021. Emotions: The unexplored fuel of fake news on social media. *J. Manag. Inf. Syst.*, 38(4):1039–1066.

Yue Huang and Lichao Sun. 2024. Fakegpt: Fake news generation, explanation and detection of large language models. *Preprint*, arXiv:2310.05046.

Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2022. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 596–605. IEEE.

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

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024b. MCFEND: A multi-source benchmark dataset for chinese fake news detection. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4018–4027. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Killian L. McLoughlin, William J. Brady, Aden Goolsbee, Ben Kaiser, Kate Klonick, and M. J. Crockett. 2024. Misinformation exploits outrage to spread online. *Science*, 386(6725):991–996.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *CoRR*, abs/1902.06673.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: multi-domain fake news detection. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3343–3347. ACM.

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, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1732–1742. ACM.

OpenAI. 2023. Chatgpt: Language model by openai. `https://openai.com/chatgpt`. Accessed: 2024-12-12.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can ai-generated text be reliably detected? *Preprint*, arXiv:2303.11156.

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, KDD 2019*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Ling Sun, Yuan Rao, Yuqian Lan, Bingcan Xia, and Yangyang Li. 2023. HG-SL: jointly learning of global and local user spreading behavior for fake news early detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 5248–5256. AAAI Press.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *CoRR*, abs/2403.18249.

Vaibhav Vaibhav, Raghuram Mandyam Annasamy, and Eduard H. 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@EMNLP 2019, Hong Kong, November 4, 2019*, pages 134–139. Association for Computational Linguistics.

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

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: generating reactions and explanations for llm-based misinformation detection. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2637–2667. Association for Computational Linguistics.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2452–2463. ACM.

Haoran Wang, Yingtong Dou, Canyu Chen, Lichao Sun, Philip S. Yu, and Kai Shu. 2023. Attacking fake news detectors via manipulating news social engagement. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3978–3986. ACM.

Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024b. Megafake: A theory-driven dataset of fake news generated by large language models. *CoRR*, abs/2408.11871.

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 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3367–3378. ACM.

Lianwei Wu, Yuan Rao, Ambreen Nazir, and Haolin Jin. 2020a. Discovering differential features: Adversarial learning for information credibility evaluation. *Inf. Sci.*, 516:453–473.

Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2020b. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3465–3476. ACM / IW3C2.

Yizhou Zhang, Defu Cao, and Yan Liu. 2022. Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

A Datasets

The details are shown in Table 5.

Politifact and **Gossipcop**: English datasets sourced from the widely used FakeNewsNet repository (Shu et al., 2020).

- **Politifact**: Contains news primarily related to political topics.
- **Gossipcop**: Includes news mainly covering entertainment topics.

Mcfend and **Weibo21**: Chinese datasets.

- **Mcfend**: Sampled from the dataset collected by (Li et al., 2024b), which aggregates data from multiple Chinese and English fact-checking websites. We selected news samples with network structures and balanced the number of true and fake news.
- **Weibo21**: A multi-domain dataset originating from Weibo, initially introduced by (Nan et al., 2021) and later supplemented with user features by (Li et al., 2024b). In this study, we utilized the enhanced version of the Weibo21.

B Baselines

News content-based detectors: detectors utilize only the text of the news article:

- **EANN** (Wang et al., 2018): Generates event invariant feature representations with adversarial training, we use the text-only version for this work.
- **SentGCN** (Vaibhav et al., 2019): Divides each news article into a graph of sentences and uses GNN for modeling.
- **BERT** (Devlin et al., 2019) and **RoBERTa** (Liu et al., 2019): Large pre-trained language models with basic settings.

Social context-enhanced detectors: detectors incorporate users' engagement information to improve detection:

- **dEFEND** (Shu et al., 2019): Develops an RNN-based sentence-comment co-attention network.
- **DualEmo** (Zhang et al., 2021): Considers emotions from publishers and users' comments, as well as the gap to improve detection.
- **BiGCN** (Bian et al., 2020): Leverages top-down and bottom-up GCN to learn the patterns of rumor propagation and dispersion respectively.
- **UPFD**: Combines news content with user preferences using graph modeling, we use two versions (**-preferences, -profile**) to show the importance of user features.

LLM-enhanced detectors: detectors utilize the inference and generation capabilities of LLMs to enhance detection:

- **ChatGPT** (chatgpt-4o-mini) (OpenAI, 2023) and **Claude** (claude-3-haiku) (Anthropic, 2024): Use zero-shot prompting to identify the veracity of news based on news content.
- **L-Defense** (Wang et al., 2024a): Divides evidence into two competing groups and asks LLMs to generate reasons for each possible veracity. In this work, we use user responses as evidence for training.
- **GenFEND** (Nan et al., 2024): Utilizes LLMs as comment generators, generating comments from potential users with diverse profiles to enhance data. We generate 15 comments for each news.

C Inferred User Intent Distribution

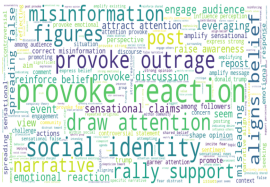
C.1 Intent Distribution

We analyzed the distribution of user intentions across four datasets, with word clouds illustrating the findings (Fig.6). The analysis reveals that, for fake news, the dominant themes are emotional responses, skepticism, and divisive topics, especially for political news, suggesting an intent to stir political debates or amplify controversial content. Conversely, real news discussions center around "inform others," "raise awareness," and "promote support", which emphasize information sharing, awareness-raising, and constructive engagement, reflecting an intent to provide accurate information and engage in rational discussions. These insights demonstrate the importance of understanding user intent in detecting and mitigating the spread of misinformation, as it reveals the underlying motivations behind user interactions.

However, for Gossipcop, user intents exhibit a certain degree of overlap, with both primarily focusing on raising questions or engaging the audience. This highlights the blurred lines between user

Table 5: Statistics of datasets (F:fake news, R:real news).

		Politifact		Gossipcop		Mcfend		Weibo21	
		F	R	F	R	F	R	F	R
News		341	240	3,430	6,903	365	200	4,488	4,640
Behaviors	post	8,354	5,988	37,728	101,456	4,949	2,752	9,208	10,127
	repost	5,587	2,967	10,324	20,190				
	comment	1,339	869	4,292	5,201	7,899	4,501	55,694	25,493
Users		18,217		46,758		15,976		92,841	



(a) Politifact-Fake

(b) Politifact-Real

(c) Gossipcop-Fake

(d) Gossipcop-Real


(e) Mcfend-Fake


(f) Mcfend-Real


(g) Weibo21-Fake


(h) Weibo21-Real

Figure 6: Word cloud of user intention distribution in true and fake news dissemination.

reactions to real and fake news in entertainment domains, and explains why the user's identity is more important in detection for this dataset.

C.2 LLM Comparison

We tried three different LLMs on two datasets to analyze whether the intent they infer is of different importance to fake detection and our framework. The results are shown in Table 6. Different LLMs exhibit varying capabilities in both Intent-only and TPB-Intent settings. Among the models, ChatGPT achieves the highest overall performance, particularly with TPB-Intent (0.955 Acc./0.953 F1 on Politifact, 0.800 Acc./0.771 F1 on Mcfend). This suggests that ChatGPT is better equipped to leverage both intent inference and TPB-guided feature aggregation for robust fake news detection. Claude underperforms relative to ChatGPT in the Intent-only setting, but the gap narrows with TPB-Intent, highlighting the effectiveness of a well-designed feature aggregation approach when individual feature is not inherently strong. QWen has consistent performance across Intent-only and TPB-Intent settings, demonstrating its superior intent reasoning capabilities. To reflect the importance of both user intent and TPB-guided aggregation, we ultimately select ChatGPT as the reasoning model.

Table 6: Performance of different LLMs in detection.

Method		Politifact		Mcfend	
		Acc.	F1	Acc.	F1
Chatgpt 4o-mini	Intent-only	0.931	0.929	0.777	0.740
	TPB-Intent	**0.955**	**0.953**	**0.800**	**0.771**
Claude 3-Haiku	Intent-only	0.912	0.909	0.790	0.770
	TPB-Intent	0.952	0.950	0.793	0.769
QWen-Turbo	Intent-only	**0.940**	**0.938**	**0.796**	**0.765**
	TPB-Intent	0.955	0.952	0.796	0.770