Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.086702

ARTICLE

Social Reaction-Aware Heterogeneous Graph Modeling for Unseen Source-Group Fake News Detection

Rongfa Chen[1,*], Liping Chen[2], Xiuzhe Meng[1] and Daniel Zeng[1,2,3]

[1]College of Management and Economics, Tianjin University, No. 92 Weijin Road, Tianjin, China
[2]School of Economics and Management, Beijing Institute of Technology, No. 5 South Zhongguancun Street, Beijing, China
[3]The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing, China

*Corresponding Author: Rongfa Chen. Email: rongfa_chen@tju.edu.cn

Received: 05 June 2026; Accepted: 07 August 2026; Published: 15 September 2026

ABSTRACT: Existing fake news detection methods largely rely on single-source datasets, leading models to overfit platform-specific features and perform poorly on heterogeneous multi-source data. Even with the emergence of Large Language Models (LLMs), our benchmarks show that general-purpose LLMs still struggle to identify deceptive intent when source-specific context is unavailable. To address unseen-source-group generalization, we propose SHIELD (Social Heterogeneous Interaction Embedding for Latent Deception). SHIELD models interaction patterns shared across sources rather than relying only on isolated text features or semantic inference. Specifically, we construct a Social Reaction-Aware Heterogeneous Interaction Graph to capture consistencies and discrepancies between news claims and user reactions, supported by constrained semantic and stylistic anchors. We introduce a Hierarchical Attentive Aggregation mechanism to learn more transferable representations from structural patterns and sentiment feedback. Empirical results on the MCFEND benchmark, where G1 denotes diverse fact-checking sources, G2 denotes translated English fact-checking sources, and G3 denotes Weibo-source news, show that SHIELD remains competitive in mixed-source detection and achieves the highest Avg.F1 among the evaluated baselines on two controlled unseen-source-group subsets with less-skewed target-label distributions. Specifically, when trained on G3 and tested on G2-Controlled and G1-Controlled target subsets, SHIELD improves Avg.F1 by 2.19 and 3.26 percentage points over the strongest trained baseline, respectively. These results suggest that structural interaction indicators, supported by training-only semantic and stylistic anchors, can improve robustness under the evaluated source-group shifts.

KEYWORDS: Fake news detection; heterogeneous graph neural networks; social reaction; source-group generalization; deception cues

1 Introduction

The proliferation of social media has made fake news a persistent challenge to public trust, and LLMs can further amplify this problem by generating convincing misinformation. At the same time, fake news detection is moving from single-platform benchmarks toward multi-source settings. Recent studies, such as MCFEND [1], show that news from social platforms, messaging applications, fact-checking agencies, and online news outlets can differ in language and context. As a result, detectors trained and evaluated within one platform or source often degrade when deployed on previously unseen sources.

Many detection methods follow a content-driven design and rely on keywords, topics, or semantic patterns. LLMs provide stronger general reasoning ability, but they still lack task-specific source context and

structured social evidence in a zero-shot setting. Our evaluations show that such models do not consistently achieve satisfactory performance, highlighting the limitations of relying only on semantic reasoning for unseen-source-group detection.

In this paper, we focus on *Unseen Source-Group Fake News Detection*, a stricter source-group generalization setting. Unlike conventional cross-domain detection, which mainly emphasizes topic or event shifts, and unlike domain adaptation, which may access target-domain samples during training, our setting assumes that the target source group is unavailable during both training and model selection. The detector must therefore learn transferable cues from observed source groups and apply them to a new group without observing target-group examples.

To address this challenge, we shift the detection focus from content alone to structured interaction analysis. Our hypothesis is that although surface-level linguistic patterns vary across platforms and sources, fake news often leaves useful signals in the interaction between claims and public reactions, such as semantic inconsistency, user skepticism, emotional mismatch, and unusual expression patterns. These claim-reaction cues are less tied to platform-specific keywords.

SHIELD does not rely on social reactions alone. It treats the claim-reaction structure as the main transferable signal and uses word and style nodes as constrained auxiliary anchors. Word nodes provide a semantic bridge between claims and reactions, while style nodes capture rule-based expression patterns such as repetition, emoticons, and POS-level cues. These components help the model determine whether reactions support, question, or conflict with the claim. To avoid target-group leakage and source-specific shortcuts, their vocabularies and corpus-level statistics are built only from the observed training groups and are frozen before validation and unseen-group testing.

We propose SHIELD (Social Heterogeneous Interaction Embedding for Latent Deception), which constructs a Social Reaction-Aware Heterogeneous Interaction Graph over news claims, user reactions, and constrained semantic/style anchors. We further design a Hierarchical Attentive Aggregation mechanism for noise-aware neighbor weighting and multi-view fusion. The main contributions are:

- We formalize Unseen Source-Group Fake News Detection as a source-group generalization task and empirically examine the limitations of semantic/content-based reasoning in this setting, using both deep learning baselines and representative LLMs as benchmarks.
- We propose the SHIELD framework, which shifts detection from linear sequence modeling toward structural dependency mining via a hierarchical attention mechanism.
- Experiments on the MCFEND benchmark show that SHIELD remains competitive with recent graph-based detectors in mixed-source detection and demonstrates improved robustness across the evaluated unseen-source-group settings.

2 Related Work

Fake news detection lies at the intersection of natural language processing (NLP) and computational social science. Existing methods can be broadly categorized into content-based analysis and structure-aware modeling. This section reviews three areas: content-based detection, GNN-based text analysis, and heterogeneous graph representation learning, with particular attention to graph-enhanced fake news detectors that are closely related to our experimental baselines.

2.1 Content-Based Fake News Detection

Content-based detection aims to distinguish deceptive content by mining linguistic indicators and semantic patterns. The field has evolved from handcrafted feature engineering to automated representation learning.

Early approaches relied on feature engineering, using manually defined indicators such as n-grams, TF-IDF, and syntactic structures with traditional classifiers such as SVMs [2,3]. To improve interpretability, researchers also used writing style and readability metrics, such as punctuation distribution and lexical diversity, to describe sensationalism and complexity in fabricated content [4–6]. Psycholinguistic features have also been used to model emotional intensity and cognitive processes in disinformation [7].

With the adoption of deep learning, end-to-end neural models have become common. CNNs and RNNs have been used to capture local semantics and temporal dependencies, respectively. More recently, pretrained language models (PLMs), particularly bidirectional Transformer-based models such as BERT, have improved performance by learning contextual representations [8,9]. Long-document variants can reduce truncation in long texts, while hybrid frameworks combine deep representations with external evidence retrieval to verify claim consistency [10,11].

Despite these advances, content-based methods still face limitations. Feature engineering relies heavily on domain expertise and scales poorly, while deep sequence models often struggle to capture long-range dependencies and remain susceptible to overfitting platform-specific semantics [12–14].

2.2 GNN-Based Text Analysis

To address the limitations of sequence models in capturing long-range semantic dependencies, Graph Neural Networks (GNNs) have been introduced to model the non-Euclidean structure of data.

In the context of text analysis, news articles are frequently transformed into graph structures. Approaches such as word-document or word-word graphs are constructed to model global word co-occurrence and shared semantics [15].

Beyond internal text structure, GNNs facilitate the integration of social context. Jointly modeling textual semantics with social network structures has proven effective in improving detection accuracy [16]. GETAE is a graph information-enhanced deep neural network ensemble architecture for fake news detection [17]. Other recent studies have attempted to fuse content graphs with propagation networks to mitigate noise in cross-platform scenarios [18]. Furthermore, dynamic weighted graphs incorporating timestamps have been explored to approximate real-world propagation mechanisms, thereby enhancing model interpretability [19].

2.3 Heterogeneous Graph Representation Learning

While homogeneous graphs capture basic topological structures, real-world scenarios involve multiple entity types (e.g., news, users, and events) and complex interactions. Heterogeneous graph representation learning addresses this complexity by distinguishing between node and edge types to learn richer semantic representations.

Heterogeneous frameworks enable the fusion of news content, user behavior, and social structures within a unified latent space. Researchers have constructed multi-level heterogeneous graphs connecting news, publishers, and social data to enhance discriminative capabilities [20]. Early studies used meta-paths for cross-type embedding, while more recent approaches apply Heterogeneous Graph Transformers to model large-scale interactions through attention mechanisms [21].

Recent graph-enhanced fake news detection methods further exploit external knowledge and multi-view consistency. For example, CONGRAT constructs heterogeneous graphs over news content, entities, topics, and external knowledge, then applies contrastive learning to align representations from multiple knowledge-augmented graph views [22]. This design improves in-domain detection by introducing complementary factual evidence from knowledge graphs. Other recent studies model publisher credibility and propagation patterns, focusing on who spreads a claim and how it spreads [23]. However, these models often depend on external knowledge resources, historical metadata, or source-specific propagation structures, which may be incomplete for unseen source groups. This motivates our focus on claim-reaction interactions: whether public responses support, question, or conflict with the claim itself.

3 Methodology

In this section, we present the SHIELD framework. We first define Unseen Source-Group Fake News Detection, then describe the construction of the Social Reaction-Aware Heterogeneous Graph, and finally present the hierarchical attentive interaction learning process.

3.1 Problem Definition

We formulate Unseen Source-Group Fake News Detection as a binary classification task under source-group generalization. Let $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ be the dataset, where each instance d_i consists of a news claim (Main Post) m_i, its associated set of social reactions (Replies) $R_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,k}\}$, a ground-truth label $y_i \in \{0, 1\}$, and a source-group identifier $s_i \in \mathcal{S}$. Here, $y_i = 0$ denotes fake news and $y_i = 1$ denotes real news. A *source group* refers to an experimental grouping of one or more origins or verification channels from which news items are collected, such as fact-checking agencies, social platforms, messaging applications, or online news outlets.

Given a set of training source groups $\mathcal{S}_{train}$ and a disjoint set of unseen test source groups $\mathcal{S}_{test}$, the key constraint is:

$$\mathcal{S}_{train} \cap \mathcal{S}_{test} = \varnothing. \tag{1}$$

The model is trained and validated only on samples from $\mathcal{S}_{train}$, and then evaluated on samples from $\mathcal{S}_{test}$. No labeled or unlabeled samples from $\mathcal{S}_{test}$ are used during training, validation, or model selection. The source-group identifier is used only to construct the evaluation split and is not used as an input feature.

This task differs from conventional cross-domain fake news detection, which mainly focuses on topic or event distribution shifts, and from cross-platform detection, which emphasizes platform-level differences such as Weibo vs. Twitter. It also differs from domain adaptation because the target source group is unavailable during training. Therefore, our task is closer to source-group domain generalization: the model must learn transferable deception cues from observed source groups and apply them to a previously unseen group.

To ensure enough structural information for graph-based modeling, we consider only instances where the number of valid social reactions k exceeds a predefined threshold (specifically, $k > 5$). The empirical choice of this threshold is examined in Section 4.2. The final dataset contains $N = 10{,}303$ refined instances. Our objective is to learn a mapping function $f_\theta : (m_i, R_i) \rightarrow \hat{y}_i$ from samples in $\mathcal{D}_{\mathcal{S}_{train}}$ that generalizes to samples in $\mathcal{D}_{\mathcal{S}_{test}}$.

Fig. 1 presents the overall workflow of SHIELD under the strict inductive unseen-source-group protocol.

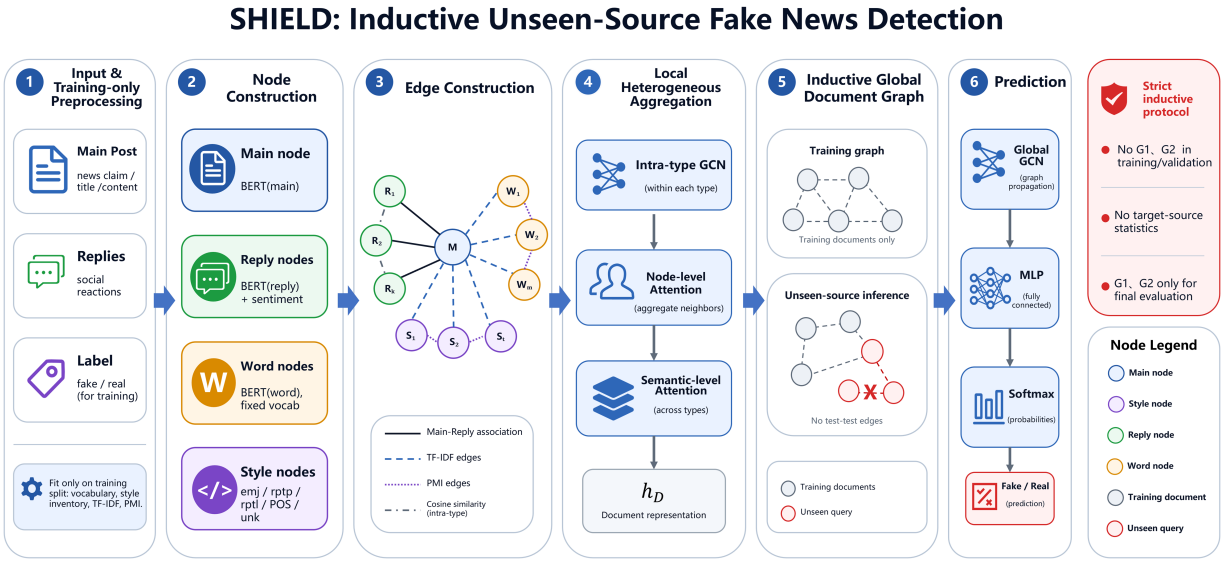

Figure 1: Overview of SHIELD under the strict inductive unseen-source-group protocol. Training-only statistics are frozen before validation and testing. Each unseen-group document is treated as an independent query node connected only to training-group documents.

3.2 Social Reaction-Aware Heterogeneous Graph Construction

To model claim-reaction interactions across source groups, we represent each news item as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

3.2.1 Training-Only Vocabulary and Statistics

To preserve the strict unseen-source-group setting, all vocabulary-level and corpus-level statistics are estimated exclusively from the training partition of the observed groups. Specifically, the word vocabulary $\mathcal{V}_w$, style vocabulary $\mathcal{V}_s$, TF-IDF statistics, and PMI-based co-occurrence matrices are constructed only from the training split of $\mathcal{D}_{\mathcal{S}_{train}}$. These statistics are frozen after training. For validation samples and unseen-group test samples, words and style markers are projected onto the fixed training vocabularies. Words absent from $\mathcal{V}_w$ are ignored, while unseen style tags are mapped to `unk`; neither type of unseen pattern is used to expand the graph. Thus, no lexical, stylistic, or co-occurrence statistics from validation samples or the unseen target group are used in graph construction. Because word and style nodes may contain source-specific lexical or stylistic patterns, SHIELD treats them as auxiliary views rather than standalone decision rules. Their contributions are adaptively weighted together with social reaction signals through the hierarchical attention mechanism.

3.2.2 Role of Word and Style Nodes in Generalization

Although word and style nodes introduce lexical and stylistic information, they are used as auxiliary anchors rather than independent source-specific predictors. Word nodes align semantic units shared by the main post and its replies, helping the model capture whether user reactions support, question, or contradict the claim. Similarly, style nodes encode rule-based and low-level expression patterns, such as emoticons, repeated tokens, repeated characters, and POS-level cues, which characterize how users react without relying on platform metadata or user identities.

As described above, the word and style representations are projected onto training-only vocabularies. They are then integrated with the main-post and reply representations through hierarchical attention, allowing SHIELD to down-weight unstable source-specific lexical and stylistic cues and emphasize more transferable claim-reaction interaction patterns.

3.2.3 Node Construction

The heterogeneous graph contains four node types: Main, Reply, Word, and Style Nodes. According to their functional roles, these nodes are organized into two categories. Main Nodes serve as central nodes, whereas Reply, Word, and Style Nodes serve as auxiliary nodes. Formally, the complete node set is defined as

$$\mathcal{V} = \mathcal{V}_m \cup \mathcal{V}_r \cup \mathcal{V}_w \cup \mathcal{V}_s,$$

where $\mathcal{V}_m$ denotes the set of central Main Nodes, while $\mathcal{V}_r$, $\mathcal{V}_w$, and $\mathcal{V}_s$ denote the sets of auxiliary Reply, Word, and Style Nodes, respectively.

BERT Feature Configuration

SHIELD represents Main, Reply, and Word Nodes using Bidirectional Encoder Representations from Transformers (BERT) [8]. Specifically, their initial features are precomputed using the BERT-wwm checkpoint released through the Chinese-BERT-wwm project (official repository: https://github.com/ymcui/Chinese-BERT-wwm) [24]. Thus, all three textual node types share a common BERT representation source and embedding space rather than using separate type-specific encoders. The resulting BERT vectors remain fixed throughout SHIELD training: they are loaded only as initial node features and are not updated through backpropagation. Consequently, SHIELD does not jointly optimize or fine-tune BERT; only the downstream graph convolution, attention, and classification parameters are trained.

1. Main Nodes ($\mathcal{V}_m$): The Discriminative Center: These nodes represent the content of the news claim. We treat the Main Node as the central hub for aggregation, combining information from reactions, semantic anchors, and style anchors. Its feature is initialized from the fixed BERT representation:

$$\mathbf{h}_{m_i}^{(0)} = \mathrm{BERT}(m_i) \in \mathbb{R}^{d_{bert}} \tag{2}$$

2. Reply Nodes ($\mathcal{V}_r$): Social Context: These nodes represent public reactions to the claim and provide sentiment and stance evidence. For the j-th reply r_j, its initial feature is:

$$\mathbf{h}_{r_j}^{(0)} = \mathrm{BERT}(r_j) \oplus \mathrm{SnowNLP}(r_j) \in \mathbb{R}^{d_{bert}+1} \tag{3}$$

Here, $\oplus$ denotes vector concatenation. SnowNLP (official repository: https://github.com/isnowfy/snownlp) is a Chinese natural language processing toolkit; $\mathrm{SnowNLP}(r_j) = s_j \in [0,1]$ denotes the sentiment probability assigned to reply r_j. Higher values indicate more positive sentiment, whereas lower values indicate more negative sentiment. This scalar provides an explicit affective cue that complements the fixed BERT representation; it is not interpreted as a direct prediction of stance or news veracity. The BERT dimensionality is $d_{bert} = 768$, so the concatenated Reply Node feature has dimension $d_{bert} + 1 = 769$.

Because the SnowNLP sentiment model was developed primarily from review-domain Chinese text, its scores may be less reliable for translated content, informal social-media replies, implicit attitudes, and sarcastic expressions. SHIELD therefore uses the score only as an auxiliary feature alongside BERT and graph-interaction evidence. Section 4.4.1 evaluates its practical contribution on G2-Controlled and G1-Controlled; direct human-annotated calibration remains outside the scope of this study.

The LLM-assisted reply selection is intended to improve the informational validity of social feedback rather than to select a particular stance. The selection prompt asks the LLM to retain replies that contain factual judgment, stance expression, logical questioning, or content-relevant evidence; it does not instruct the LLM to prefer replies that support or oppose the news claim. Replies are mainly removed when they are empty, purely emotive, repetitive, unrelated, or too short to provide useful claim-reaction evidence. Appendix A.1 provides the detailed prompt, fallback rule, and rationale for the 10-reply cap.

3. Word Nodes ($\mathcal{V}_w$): Semantic Units: To capture fine-grained semantic cues, we construct a fixed vocabulary from the main posts and replies in the training groups. The resulting word nodes align semantic units shared by claims and reactions and are initialized as $\mathbf{h}_{w_k}^{(0)} = \mathrm{BERT}(w_k)$.
4. Style Nodes ($\mathcal{V}_s$): Latent Style: To capture non-semantic stylistic features, we apply a rule-based style extraction procedure to training-group replies. Each reply is converted into a sequence of style tags rather than a single label. As shown in Table 1, the rules cover emoticon/symbol patterns, adjacent token repetition, repeated characters within a token, and POS tags. Chinese POS tags are generated by `jieba.posseg`; tokens not covered by the Chinese tagger use an NLTK POS tagger as a fallback, and unresolved tokens are assigned `unk`. The style vocabulary $\mathcal{V}_s$ is collected automatically from the training split and then fixed. During validation and unseen-group testing, extracted style tags are projected onto this fixed vocabulary, while unseen tags are mapped to `unk` and never used to expand $\mathcal{V}_s$. Because style symbols do not have rich semantic context, we initialize their features using one-hot encoding, represented by an identity matrix $\mathbf{I} \in \mathbb{R}^{|\mathcal{V}_s| \times |\mathcal{V}_s|}$, and learn their embeddings during training.

Table 1: Rule-based style tag extraction.

Style Type	Tag	Rule
Emoticon/Symbol	`emj`	Common emoticons or bracketed expressions
Adjacent repetition	`rptp`	Current token equals the next token
Character repetition	`rptl`	Consecutive repeated characters within a token
Part-of-speech	POS tag	Assigned by jieba.posseg
Unknown	`unk`	No valid style or POS tag is obtained

Note: The table lists the rule-based style tags used to construct style nodes.

3.2.4 Edge Construction

We define edges to capture interactions centered around the Main Node:

- Main-Reply Edges ($\mathbf{A}_{mr}$): These edges connect the main post m_i to its replies r_j, allowing the Main Node to aggregate social conflict and sentiment variation.
- Main-Semantic Edges ($\mathbf{A}_{mw}$) & Main-Style Edges ($\mathbf{A}_{ms}$): These edges connect the Main Node to keywords and style symbols in the fixed training vocabulary based on TF-IDF statistics fitted to the training-group corpus. They provide constrained semantic and stylistic anchors for interpreting claim-reaction relationships without using target-group statistics.
- Semantic Co-occurrence Edges ($\mathbf{A}_{ww}$) & Style Co-occurrence Edges ($\mathbf{A}_{ss}$): To capture co-occurrence patterns in the training groups (e.g., word-word co-occurrence and style-style association), we compute weights using Pointwise Mutual Information (PMI) on the training-group corpus only:

$$A_{ww_{ij}} = \max(0, \mathrm{PMI}(w_i, w_j))$$
$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \tag{4}$$

These edges form homogeneous subgraphs that help the model learn semantic and stylistic representations for source-group generalization.

3.3 Hierarchical Attentive Interaction Learning

After constructing the heterogeneous graph, SHIELD uses a hierarchical attentive mechanism to aggregate information dynamically. This process consists of three steps.

3.3.1 Step 1: Intra-Graph Contextualization

Before modeling heterogeneous interactions, we propagate information within the homogeneous subgraphs to obtain local context representations. For each node type $\tau \in \{r, w, s\}$, we use a two-layer GCN to update its features:

$$\mathbf{H}_\tau = GCN_{\tau,2}\left(A_{\tau\tau}, ReLU\left(GCN_{\tau,1}(A_{\tau\tau}, X_\tau)W_{\tau,1}\right)\right)W_{\tau,2} \tag{5}$$

where $\mathbf{A}_{\tau\tau}$ corresponds to the homogeneous adjacency matrices defined above ($\mathbf{A}_{ww}$, $\mathbf{A}_{ss}$). Specifically, for Reply Nodes, we dynamically construct $\mathbf{A}_{rr}$ based on cosine similarity to capture associations between similar replies.

3.3.2 Step 2: Node-Level Attentive Aggregation

In this step, the Main Node $v_i \in \mathcal{V}_m$ aggregates information from different types of neighbors (Reply, Word, Style). Because these neighbors may carry different levels of evidence, we introduce type-specific attention mechanisms. For a Main Node v_i and its neighbor $v_j \in \mathcal{N}_i^\tau$ of type τ, the attention coefficient α_{ij}^τ is computed as:

$$\begin{aligned} e_{ij}^\tau &= \mathrm{ReLU}\left(\mathbf{a}_\tau^T[\mathbf{W}_\tau\mathbf{h}_i \| \mathbf{W}_\tau\mathbf{h}_j]\right) \\ \alpha_{ij}^\tau &= \frac{\exp(e_{ij}^\tau)}{\sum_{k\in\mathcal{N}_i^\tau}\exp(e_{ik}^\tau)} \end{aligned} \tag{6}$$

Using the normalized weights, we obtain the view-specific aggregated representation $\mathbf{z}_i^\tau$ of the Main Node.

3.3.3 Step 3: Semantic-Level Attentive Fusion

After node-level aggregation, each Main Node v_i has three view representations, $\{\mathbf{z}_i^{reply}, \mathbf{z}_i^{word}, \mathbf{z}_i^{style}\}$, together with its own representation $\mathbf{z}_i^{self}$. To weigh the contributions of structural conflict, textual semantics, and stylistic patterns, we use multi-head self-attention for feature fusion:

$$\begin{aligned} \mathbf{Z}_i &= \mathrm{Stack}(\mathbf{z}_i^{self}, \mathbf{z}_i^{reply}, \mathbf{z}_i^{word}, \mathbf{z}_i^{style}) \\ \mathbf{Z}_i^{fused} &= \mathrm{LayerNorm}(\mathbf{Z}_i + \mathrm{MultiHead}(\mathbf{Z}_i)) \end{aligned} \tag{7}$$

Finally, mean pooling produces the higher-order representation $\mathbf{h}_i^{final}$ of the Main Node.

3.4 Global Prediction

To infer the authenticity of each news item and capture document-level associations, we further construct an inductive global document graph over training-group documents. We calculate the news document

representation $\mathbf{h}_D^{final}$ from the fused representation of its Main Node and construct a training adjacency matrix $\mathbf{A}_{global}^{train}$ based only on pairwise similarities among documents from the training partition of $\mathcal{S}_{train}$. The representation incorporates the fixed BERT feature and the Reply, Word, and Style evidence aggregated by the hierarchical interaction module. Pairwise edge weights are computed using cosine similarity between these fused document representations. Negative similarities are clipped to zero, and the resulting nonnegative similarity matrix is normalized row-wise so that the positive outgoing weights in each nonzero row sum to one. The reported experiments use all such soft similarity weights and do not apply Top-k selection or threshold-based pruning. Self-loops are then added, followed by the standard symmetric adjacency normalization used by the global GCN. No validation or test documents are included when constructing the training graph.

During validation and inference, each held-out document is treated as an independent query node. Its fused representation is computed using the fixed BERT feature and the trained downstream SHIELD modules, compared with all training-document representations, and connected to the training graph through the same nonnegative, row-normalized soft similarity weights. Importantly, no edges are constructed among held-out documents, and the document similarity matrix is never computed over the entire validation set or unseen-group test set. Therefore, SHIELD does not access the distributional structure of the target group during training or prediction.

The final prediction is produced by a global GCN layer and a softmax classifier:

$$\hat{y} = \text{Softmax}(\text{MLP}(\text{GCN}_{final}(\mathbf{A}_{global}^{*}, \mathbf{H}_D^{final}))) \tag{8}$$

where $\mathbf{A}_{global}^{*}$ denotes $\mathbf{A}_{global}^{train}$ during training and the query-augmented training graph during inference. This design allows the model to leverage document-level relations learned from observed groups while maintaining a strictly inductive unseen-source-group protocol.

4 Experiments

In this section, we evaluate SHIELD on the multi-source MCFEND benchmark to answer the following research questions (RQs):

- RQ1 (Generalization): How does SHIELD perform under mixed-source and controlled unseen-source settings compared with content-based and graph-based baselines?
- RQ2 (Ablation): How do social-reaction, semantic, stylistic, and sentiment components affect unseen-source-group performance?

4.1 Experimental Setup

4.1.1 Dataset and Source Grouping

We evaluate our method on MCFEND [1], a Chinese benchmark for multi-source fake news detection. The raw dataset contains 30,676 news items from multiple platforms.

We applied a multi-stage filtering pipeline before model evaluation:

1. **Label Cleaning**: We excluded samples labeled as "undetermined" (approximately 0.6% of the raw data).
2. **LLM-assisted Reply Selection**: We employed an LLM-assisted selection module to retain representative replies and reduce redundant or irrelevant comments in the reply sets. The detailed selection criteria and fallback strategy are provided in Appendix A.1.

3. **Interaction Sparsity Filtering**: As defined in Section 3.1, we retained only news items with more than five valid replies after reply selection so that each graph contains sufficient reaction information. The effect of alternative thresholds is evaluated in Section 4.2.

After filtering, the final experimental dataset contains 6427 fake news items and 3876 real news items. To evaluate source-group generalization, we categorize the data into three groups:

- G1 (Diverse Sources): Mixed data from 9 different fact-checking agencies (including portals and social media). It consists of 2042 samples, Real: 112 (5.48%); Fake: 1930 (94.52%).
- G2 (Translated Fact-Checking Sources): News translated from English fact-checking agencies (e.g., PolitiFact, GossipCop), with distinct linguistic styles and topic distributions. It consists of 640 samples, Real: 75 (11.72%); Fake: 565 (88.28%).
- G3 (Single Social Source): Data solely from Weibo. This is the largest group with 7621 samples, showing a relatively balanced distribution, Real: 3689 (48.41%); Fake: 3932 (51.59%).

The label and source-group distributions of the filtered dataset remain close to those of the raw dataset, with deviations within 3 percentage points.

4.1.2 Evaluation Protocol

We consider two evaluation settings. First, *Multi-Source Detection* randomly splits samples from all source groups (G1 + G2 + G3) into training, validation, and test sets with an 8:1:1 ratio. In this setting, the training and test data share the same source coverage, and the goal is to evaluate overall detection performance under a mixed-source distribution.

Second, *Controlled Unseen Source-Group Detection* evaluates source-group generalization while reducing the effect of severe target-label skew. In this setting, all evaluated supervised models are trained and validated exclusively on G3 and are then tested on two held-out target subsets with a 1:2 Real-to-Fake ratio, denoted G2-Controlled and G1-Controlled. The controlled subsets are used only for final evaluation and are never used for training, validation, model selection, graph construction, or corpus-level statistic estimation.

Dataset Split Details

To avoid ambiguity, Table 2 summarizes the exact source usage in the main experimental settings. In the multi-source detection setting, samples from G1, G2, and G3 are randomly split into training, validation, and test sets following the 8:1:1 ratio. This setting evaluates conventional mixed-source detection, where all source groups are represented in model development and testing. In the controlled unseen-source setting, only G3 samples are split into training and validation sets with an 88:12 ratio, while G2-Controlled and G1-Controlled are held out entirely as target subsets for final testing.

Table 2: Source usage under different evaluation protocols.

Setting	Training Group	Validation Group	Test Group
Multi-Source Detection	G1 + G2 + G3 (80%)	G1 + G2 + G3 (10%)	G1 + G2 + G3 (10%)
Controlled Unseen Target I	G3 (88%)	G3 (12%)	G2-Controlled (225 samples)
Controlled Unseen Target II	G3 (88%)	G3 (12%)	G1-Controlled (336 samples)

Note: Percentages indicate the split proportions applied within the corresponding source groups. The controlled target subsets are used only for final testing.

Leakage Prevention

For all unseen-source-group experiments, the target subset is excluded from model training, validation, hyperparameter selection, graph construction, and statistical estimation. The training-only construction rules described in Section 3 are enforced before evaluating either G2-Controlled or G1-Controlled as an unseen target. During testing, each target sample is projected onto the frozen training vocabularies and connected only to training-group documents in the inductive global graph. None of the controlled target samples is used for model development or corpus-level estimation.

4.1.3 Implementation Details

The fixed 768-dimensional BERT vectors described in the Node Construction subsection are loaded before graph training and are not included in the optimization process. The learning rate of 1×10^{-3} applies only to the trainable downstream SHIELD parameters; no gradients are propagated to BERT representations.

4.1.4 Baselines and Metrics

To evaluate the performance of SHIELD, we compare it against representative baselines from four categories. For fairness, we distinguish the common evaluation protocol from model-specific hyperparameter choices. All supervised trainable models use the same data splits, leakage-prevention rules, early-stopping principle, and five-run reporting protocol; the LLM baselines follow the separate zero-shot protocol described in Appendix A.2. For baselines evaluated in the MCFEND paper, we follow the public hyperparameter settings reported in that work [1]. For graph-based baselines not covered by MCFEND, including GETAE and CONGRAT, we use MCFEND-compatible shared training controls while retaining architecture-specific settings required by each model. SHIELD also adopts several common settings from MCFEND rather than using a uniquely larger tuning budget. Final configurations for supervised models are selected according to validation Avg.F1, and the held-out test subsets are never used for hyperparameter selection. We acknowledge that a broader hyperparameter search may further improve individual baselines; therefore, the comparison should be interpreted under a standardized and literature-guided tuning budget rather than an exhaustive search over all possible configurations.

Unless otherwise specified, all reported improvements and drops are absolute differences measured in percentage points (pp), rather than relative percentage changes. All experiments are repeated over five independent runs using different random seeds. Unless otherwise specified, each metric is reported as the mean ± standard deviation across these runs. Accordingly, Avg.A, Avg.P, Avg.R, and Avg.F1 denote the five-run mean Accuracy, Precision, Recall, and F1 score, respectively, with the corresponding standard deviation reported in the same table entry.

Because several source groups and target subsets exhibit label imbalance, we treat Avg.F1 as the primary metric for model comparison and robustness interpretation. Avg.A is retained as a supplementary metric, while Avg.P and Avg.R are used to clarify whether performance changes are precision- or recall-oriented.

- **Content-based:** BERT and RoBERTa. These models were trained end-to-end with a learning rate of 2×10^{-5}.
- **Social Context-based:** BERT-EMO. This model was trained end-to-end with a learning rate of 1×10^{-3}.
- **Graph-based:** TextGCN, **HGT** (Heterogeneous Graph Transformer), GETAE, and CONGRAT. TextGCN is a classic homogeneous graph neural network, while HGT serves as a strong generic heterogeneous graph baseline. GETAE is a graph information-enhanced deep neural network ensemble architecture, and CONGRAT introduces contrastive multi-knowledge graph learning for fake news detection. For graph-based baselines without public MCFEND-specific configurations, we use

MCFEND-compatible shared training controls (Learning Rate: 1e–3, Hidden Size: 400, Dropout: 0.7) and keep architecture-specific components unchanged.

- **Large Language Models (LLMs):** To evaluate zero-shot generative baselines, we include Qwen3.5-9B, GLM-Z1-9B, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-32B. We use a task-specific zero-shot prompt that asks the models to judge authenticity based on source reliability, objectivity, emotional language, and logical consistency. The temperature is set to 0.1 for all LLMs. Additional evaluation details are reported in Appendix A.2.

4.2 Sensitivity to the Interaction Sparsity Threshold

To examine whether the choice of $k > 5$ is supported empirically, we construct five dataset variants by requiring more than 1, 3, 5, 7, or 10 candidate valid replies before applying the 10-reply selection cap. The subsequent LLM-assisted selection procedure is kept unchanged across variants. GETAE and CONGRAT are selected as two strong graph-enhanced baselines for this threshold-selection experiment. The same training and evaluation procedure is used across all five variants, and each result is reported as the mean ± standard deviation over five independent runs.

Table 3 and Fig. 2 show a consistent non-monotonic pattern. Both baselines achieve their best performance at $k > 5$. GETAE reaches an F1 score of 0.9040, improving by 4.92 and 4.56 percentage points over the $k > 1$ and $k > 3$ settings, respectively. CONGRAT reaches an F1 score of 0.9105, corresponding to improvements of 5.48 and 3.59 percentage points. Increasing the threshold to $k > 7$ or $k > 10$ reduces performance, indicating that progressively stricter filtering does not continuously improve graph learning. We therefore use $k > 5$ as an empirical trade-off between obtaining sufficient reaction evidence and avoiding unnecessarily restrictive filtering. Nevertheless, this analysis does not remove the deployment limitation for documents with five or fewer valid replies, which remains outside the scope of the current evaluation.

Table 3: Performance under different minimum-reply thresholds.

Model	Threshold	Accuracy	Precision	Recall	F1
GETAE	$k > 1$	0.8880 ± 0.0105	0.8734 ± 0.0133	0.8399 ± 0.0112	0.8548 ± 0.0118
GETAE	$k > 3$	0.8940 ± 0.0080	0.8652 ± 0.0095	0.8526 ± 0.0087	0.8584 ± 0.0081
GETAE	$k > 5$	**0.9026 ± 0.0041**	**0.9037 ± 0.0049**	**0.9043 ± 0.0033**	**0.9040 ± 0.0038**
GETAE	$k > 7$	0.8656 ± 0.0060	0.8484 ± 0.0108	0.8009 ± 0.0085	0.8206 ± 0.0075
GETAE	$k > 10$	0.8786 ± 0.0113	0.8612 ± 0.0217	0.8418 ± 0.0127	0.8504 ± 0.0156
CONGRAT	$k > 1$	0.8885 ± 0.0044	0.8761 ± 0.0120	0.8390 ± 0.0048	0.8557 ± 0.0075
CONGRAT	$k > 3$	0.8997 ± 0.0065	0.8869 ± 0.0118	0.8641 ± 0.0090	0.8746 ± 0.0067
CONGRAT	$k > 5$	**0.9066 ± 0.0044**	**0.9134 ± 0.0022**	**0.9079 ± 0.0025**	**0.9105 ± 0.0022**
CONGRAT	$k > 7$	0.8685 ± 0.0053	0.8586 ± 0.0089	0.8128 ± 0.0054	0.8324 ± 0.0068
CONGRAT	$k > 10$	0.8858 ± 0.0076	0.8793 ± 0.0061	0.8493 ± 0.0162	0.8623 ± 0.0101

Note: Each entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result for each model and metric.

4.3 Performance Comparison (RQ1)

4.3.1 Multi-Source Detection (All Groups Mixed)

First, we evaluate performance when training and testing on mixed data from all source groups (G1 + G2 + G3). The results are shown in Table 4.

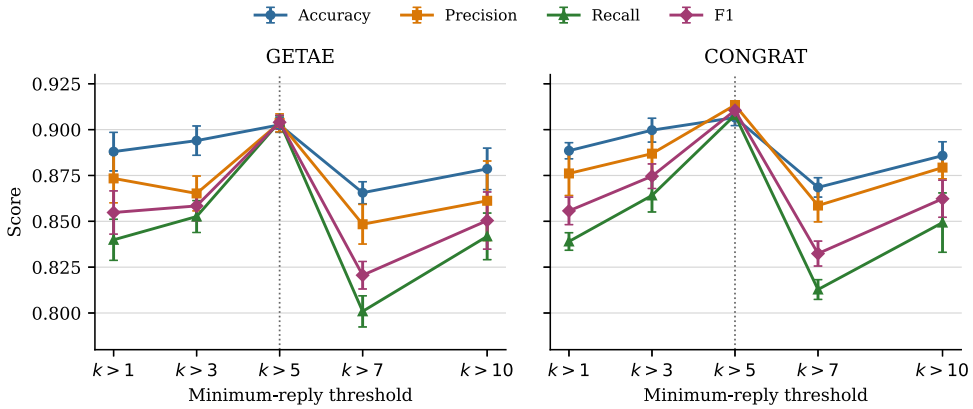

Figure 2: Sensitivity of GETAE and CONGRAT to the minimum-reply threshold. Error bars indicate the standard deviation over five independent runs.

Table 4: Performance on multi-source detection (all groups mixed).

Model	Avg.A	Avg.P	Avg.R	Avg.F1
LLM Baselines (Zero-shot)				
Qwen3.5-9B	0.6384 ± 0.0064	0.4030 ± 0.0063	0.3818 ± 0.0094	0.3921 ± 0.0059
GLM-Z1-9B	0.5005 ± 0.0021	0.5066 ± 0.0048	0.4981 ± 0.0103	0.5023 ± 0.0058
DeepSeek-R1-Distill-Llama-8B	0.4223 ± 0.0072	0.4254 ± 0.0061	0.4232 ± 0.0112	0.4242 ± 0.0066
DeepSeek-R1-Distill-Qwen-32B	0.6367 ± 0.0102	0.5139 ± 0.0077	0.4080 ± 0.0084	0.4548 ± 0.0072
Trained Baselines				
TextGCN	0.7109 ± 0.0103	0.5068 ± 0.0172	0.4443 ± 0.0166	0.4435 ± 0.0186
HGT	0.8359 ± 0.0026	0.8270 ± 0.0023	0.8226 ± 0.0037	0.8247 ± 0.0031
GETAE	0.9026 ± 0.0041	0.9037 ± 0.0049	0.9043 ± 0.0033	0.9040 ± 0.0038
CONGRAT	**0.9066 ± 0.0044**	0.9134 ± 0.0022	**0.9079 ± 0.0025**	**0.9105 ± 0.0022**
BERT	0.7466 ± 0.0051	0.8344 ± 0.0100	0.6962 ± 0.0108	0.7102 ± 0.0081
BERT-EMO	0.7591 ± 0.0055	0.8341 ± 0.0061	0.7123 ± 0.0102	0.7231 ± 0.0103
RoBERTa	0.7248 ± 0.0047	0.7609 ± 0.0035	0.6871 ± 0.0106	0.6932 ± 0.0082
SHIELD (Ours)	0.9005 ± 0.0053	**0.9177 ± 0.0037**	0.8796 ± 0.0076	0.8983 ± 0.0057

Note: Each entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result in each column. Avg.A, Avg.P, Avg.R, and Avg.F1 denote the five-run mean Accuracy, Precision, Recall, and F1 score, respectively.

As shown in Table 4, when training and test data share the same source coverage, graph-enhanced methods generally outperform content-based baselines. CONGRAT achieves the highest Avg.F1 (0.9105 ± 0.0022), indicating that contrastive multi-knowledge graph learning is effective when source distributions overlap. GETAE also performs strongly (Avg.F1: 0.9040 ± 0.0038). SHIELD reaches an Avg.F1 of 0.8983 ± 0.0057 while achieving the highest Avg.P (0.9177 ± 0.0037). This result shows that SHIELD retains competitive mixed-source performance, although its design is aimed at unseen-source-group generalization.

In the zero-shot LLM evaluation, each model receives the news text and available replies within its supported context window, as described in Appendix A.2. The best LLM Avg.F1 is achieved by GLM-Z1-9B (0.5023 ± 0.0058), which remains substantially below the strongest trained graph baselines. These results

suggest that simply providing longer serialized textual input is insufficient to match supervised graph-based detectors. However, this comparison should not be interpreted as a direct measurement of intrinsic reasoning ability: LLMs process serialized text in a zero-shot manner, whereas graph-based supervised models explicitly learn structured reply interactions and corpus-level relations.

4.3.2 Controlled Unseen Source-Group Detection

The full G2 group contains only 640 samples and is strongly skewed toward fake news. To reduce the influence of this label skew and to evaluate more than one unseen target group, we use two class-controlled target subsets as the main unseen-source evaluation. G2-Controlled retains all 75 real-news samples from G2 and randomly samples 150 fake-news samples, yielding 225 test instances with a 1:2 Real-to-Fake ratio. G1-Controlled retains all 112 real-news samples from G1 and randomly samples 224 fake-news samples, yielding 336 test instances with the same ratio. In both settings, all evaluated supervised models are trained and validated exclusively on G3, and the controlled target subset is used only for final testing. Results are reported as mean ± standard deviation over five independent runs. The controlled-target results are reported in Tables 5 and 6.

Table 5: Performance under G3→G2-Controlled.

Model	Avg.A	Avg.P	Avg.R	Avg.F1
TextGCN	0.6370 ± 0.1298	0.5193 ± 0.0036	0.5373 ± 0.0147	0.4728 ± 0.0431
HGT	0.5405 ± 0.0068	0.5537 ± 0.0064	**0.5586 ± 0.0070**	0.5349 ± 0.0067
GETAE	0.4847 ± 0.0252	0.5324 ± 0.0246	0.5313 ± 0.0243	0.4846 ± 0.0238
CONGRAT	0.4438 ± 0.0231	0.5345 ± 0.0082	0.5279 ± 0.0099	0.4371 ± 0.0287
BERT	0.5581 ± 0.0106	0.5512 ± 0.0092	0.5489 ± 0.0264	0.5489 ± 0.0120
BERT-EMO	0.5516 ± 0.0127	0.5486 ± 0.0116	0.5534 ± 0.0158	0.5499 ± 0.0082
RoBERTa	0.5739 ± 0.0056	0.5507 ± 0.0059	0.5550 ± 0.0095	0.5525 ± 0.0076
SHIELD (Ours)	**0.7074 ± 0.0282**	**0.6491 ± 0.0032**	0.5152 ± 0.0155	**0.5744 ± 0.0105**

Note: Each entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result in each column.

Table 6: Performance under G3→G1-Controlled.

Model	Avg.A	Avg.P	Avg.R	Avg.F1
TextGCN	0.4798 ± 0.0140	0.5312 ± 0.0170	0.5313 ± 0.0158	0.4783 ± 0.0125
HGT	**0.6220 ± 0.0075**	0.5643 ± 0.0114	0.5603 ± 0.0124	0.5609 ± 0.0127
GETAE	0.5649 ± 0.0186	0.5415 ± 0.0158	0.5460 ± 0.0163	0.5391 ± 0.0167
CONGRAT	0.5685 ± 0.0153	0.5611 ± 0.0077	0.5688 ± 0.0087	0.5530 ± 0.0108
BERT	0.6214 ± 0.0099	0.4628 ± 0.0313	0.4884 ± 0.0039	0.4321 ± 0.0117
BERT-EMO	0.6113 ± 0.0076	0.4546 ± 0.0276	0.4866 ± 0.0089	0.4326 ± 0.0168
RoBERTa	0.6167 ± 0.0093	0.4928 ± 0.0052	0.4978 ± 0.0020	0.4575 ± 0.0036
SHIELD (Ours)	0.6203 ± 0.0174	**0.5936 ± 0.0155**	**0.6018 ± 0.0166**	**0.5935 ± 0.0165**

Note: Each entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result in each column.

On G2-Controlled, SHIELD achieves the highest Avg.F1 (0.5744 ± 0.0105), improving over RoBERTa, the strongest baseline on this metric (0.5525 ± 0.0076), by 2.19 percentage points. It also obtains the highest Avg.P (0.6491 ± 0.0032) and Avg.A (0.7074 ± 0.0282), although HGT obtains the highest Avg.R

(0.5586 ± 0.0070). On G1-Controlled, SHIELD again achieves the highest Avg.F1 (0.5935 ± 0.0165), improving by 3.26 percentage points over HGT (0.5609 ± 0.0127). It also achieves the highest Avg.P (0.5936 ± 0.0155) and Avg.R (0.6018 ± 0.0166), while its Avg.A (0.6203 ± 0.0174) is close to, but slightly below, HGT (0.6220 ± 0.0075). These results show that SHIELD's F1 advantage persists after reducing target-label skew and when changing the controlled unseen target from G2 to G1. However, G2-Controlled and G1-Controlled remain relatively small, so they provide robustness evidence within the evaluated MCFEND setting rather than establishing unrestricted cross-source generalization.

4.4 Ablation Study (RQ2)

To align the component analysis with the controlled unseen-source evaluation, we conduct ablations on G2-Controlled and G1-Controlled, as defined in Section 4.3.2. In both settings, SHIELD is trained and validated exclusively on G3, and the controlled target subset is used only for final testing. This design examines whether Reply, Word, and Style Nodes provide consistent contributions across different unseen target groups under less-skewed target distributions. The corresponding ablation results are reported in Tables 7 and 8.

Table 7: Ablation study of SHIELD under G3→G2-Controlled.

Model Variant	Avg.A	Avg.P	Avg.R	Avg.F1	F1 Drop (pp)
SHIELD (Full)	**0.7074 ± 0.0282**	**0.6491 ± 0.0032**	0.5152 ± 0.0155	**0.5744 ± 0.0105**	–
w/o Style	0.5628 ± 0.0179	0.5964 ± 0.0187	**0.6018 ± 0.0196**	0.5614 ± 0.0179	↓ 1.30 pp
w/o Word	0.5563 ± 0.0201	0.5814 ± 0.0139	0.5875 ± 0.0159	0.5537 ± 0.0190	↓ 2.07 pp
w/o Reply	0.5218 ± 0.0157	0.5817 ± 0.0064	0.5796 ± 0.0085	0.5213 ± 0.0167	↓ 5.31 pp
w/o Word & Reply	0.5200 ± 0.0099	0.5680 ± 0.0103	0.5689 ± 0.0105	0.5199 ± 0.0099	↓ 5.45 pp
w/o Style & Reply	0.5302 ± 0.0184	0.5812 ± 0.0200	0.5817 ± 0.0200	0.5301 ± 0.0185	↓ 4.43 pp
w/o Style & Word	0.5553 ± 0.0201	0.5913 ± 0.0152	0.5960 ± 0.0170	0.5543 ± 0.0193	↓ 2.01 pp
Main Only (No Graph)	0.4995 ± 0.1262	0.3908 ± 0.1833	0.5235 ± 0.0513	0.4223 ± 0.1435	**↓ 15.21 pp**

Note: Each metric entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result in each metric column or the largest mean F1 drop. Avg.A, Avg.P, Avg.R, and Avg.F1 denote five-run means; pp denotes percentage points, and F1 Drop is calculated from the corresponding means.

Table 8: Ablation study of SHIELD under G3→G1-Controlled.

Model Variant	Avg.A	Avg.P	Avg.R	Avg.F1	F1 Drop (pp)
SHIELD (Full)	**0.6203 ± 0.0174**	**0.5936 ± 0.0155**	**0.6018 ± 0.0166**	**0.5935 ± 0.0165**	–
w/o Style	0.6190 ± 0.0122	0.5849 ± 0.0073	0.5897 ± 0.0090	0.5849 ± 0.0088	↓ 0.86 pp
w/o Word	0.6167 ± 0.0100	0.5885 ± 0.0099	0.5960 ± 0.0107	0.5887 ± 0.0102	↓ 0.48 pp
w/o Reply	0.6101 ± 0.0130	0.5790 ± 0.0115	0.5848 ± 0.0122	0.5791 ± 0.0121	↓ 1.44 pp
w/o Word & Reply	0.5935 ± 0.0048	0.5410 ± 0.0080	0.5733 ± 0.0098	0.5566 ± 0.0085	↓ 3.69 pp
w/o Style & Reply	0.5786 ± 0.0098	0.5251 ± 0.0096	0.5600 ± 0.0108	0.5419 ± 0.0101	↓ 5.16 pp
w/o Style & Word	0.6083 ± 0.0165	0.5572 ± 0.0125	0.5853 ± 0.0128	0.5709 ± 0.0139	↓ 2.26 pp
Main Only (No Graph)	0.5137 ± 0.1480	0.3897 ± 0.1829	0.5112 ± 0.0106	0.3986 ± 0.1214	**↓ 19.49 pp**

Note: Each metric entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result in each metric column or the largest mean F1 drop. Avg.A, Avg.P, Avg.R, and Avg.F1 denote five-run means; pp denotes percentage points, and F1 Drop is calculated from the corresponding means.

Fig. 3 visualizes the Avg.F1 degradation caused by removing individual or paired SHIELD components on the two controlled target subsets.

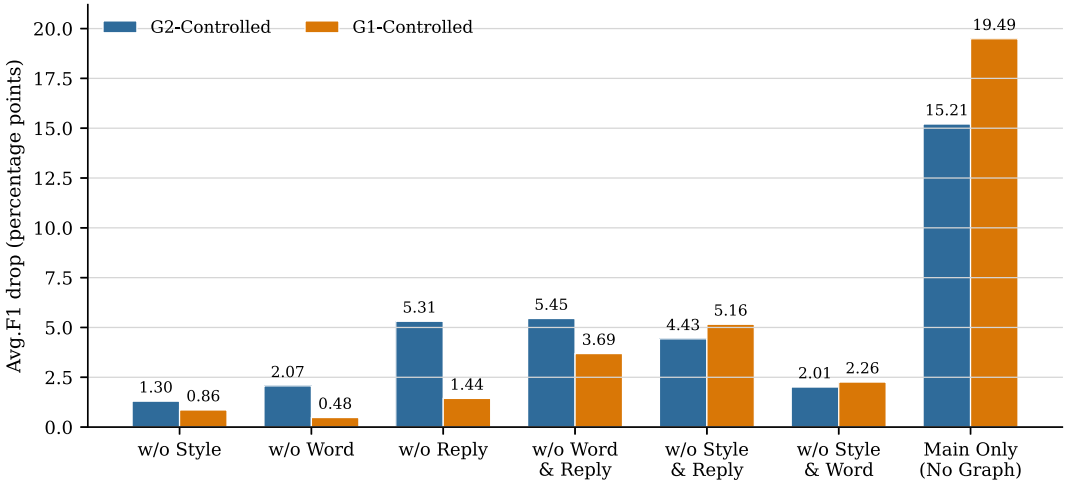

Figure 3: Avg.F1 drops of SHIELD variants on G2-Controlled and G1-Controlled.

The controlled-target ablation results show the following patterns:

1. Social Reaction: Removing the Reply Node causes the largest single-component Avg.F1 drop on both controlled subsets: 5.31 percentage points on G2-Controlled and 1.44 percentage points on G1-Controlled. This pattern indicates that claim-reaction interactions provide useful transfer evidence beyond the main post representation.
2. Word and Style Nodes: Removing Word Nodes or Style Nodes also reduces Avg.F1 on both target subsets, although the magnitude varies by target group. On G2-Controlled, the drops are 2.07 and 1.30 percentage points, respectively. On G1-Controlled, the drops are smaller, at 0.48 and 0.86 percentage points. These results suggest that semantic and stylistic anchors provide complementary cues, but their contribution depends on the target group's language and source characteristics.
3. Paired Components and Content-Only Signals: Paired removals generally cause larger degradation than most single-component removals, showing that the components are complementary rather than redundant. The Main Only (No Graph) variant is the weakest and most variable setting, with Avg.F1 dropping by 15.21 percentage points on G2-Controlled and 19.49 percentage points on G1-Controlled. This confirms that content-only representations are less reliable under controlled unseen-target evaluation, while heterogeneous interaction modeling improves robustness.

Overall, the two controlled-target ablations provide additional evidence that SHIELD's component contributions are not limited to a single controlled target subset. The direction of contribution is consistent across G2-Controlled and G1-Controlled, although the exact magnitude varies with the target group. We therefore interpret the ablation results as robustness evidence within the evaluated MCFEND source groups rather than as proof of unrestricted cross-source generalization.

4.4.1 Effect of the SnowNLP Sentiment Feature

To examine whether the auxiliary SnowNLP score contributes under different unseen target groups, we compare the full model with a variant that removes the scalar sentiment feature from every Reply Node while retaining the fixed BERT representation and all other components. We evaluate both variants

on G2-Controlled and G1-Controlled, as defined in Section 4.3.2; both variants are trained and validated exclusively on G3. Results are reported as mean ± standard deviation over five independent runs.

As shown in Table 9, including the SnowNLP score increases Avg.F1 by 2.90 percentage points on G2-Controlled and by 0.83 percentage points on G1-Controlled. The full model also improves Avg.P by 11.75 and 3.40 percentage points, respectively, although the variant without SnowNLP obtains higher Avg.R on both target subsets. The sentiment score therefore appears to provide a useful precision-oriented auxiliary cue, particularly for G2-Controlled, but it does not improve every metric uniformly. This ablation evaluates downstream utility rather than probability calibration; a human-annotated cross-source sentiment benchmark would still be needed to assess calibration directly.

Table 9: Ablation of the SnowNLP sentiment feature on G2-Controlled and G1-Controlled.

Target Subset	Variant	Avg.A	Avg.P	Avg.R	Avg.F1
G2-Controlled (1:2)	Full	**0.7074 ± 0.0282**	**0.6491 ± 0.0032**	0.5152 ± 0.0155	**0.5744 ± 0.0105**
G2-Controlled (1:2)	w/o SnowNLP	0.5846 ± 0.0121	0.5316 ± 0.0119	**0.5600 ± 0.0105**	0.5454 ± 0.0157
G1-Controlled (1:2)	Full	**0.6203 ± 0.0174**	**0.5936 ± 0.0155**	0.6018 ± 0.0166	**0.5935 ± 0.0165**
G1-Controlled (1:2)	w/o SnowNLP	0.6131 ± 0.0054	0.5596 ± 0.0051	**0.6133 ± 0.0105**	0.5852 ± 0.0075

Note: Each entry reports the mean ± standard deviation over five independent runs. Bold values indicate the best mean result within each controlled target subset.

5 Conclusion and Future Work

In this paper, we proposed SHIELD for unseen-source-group fake news detection. SHIELD models claim-reaction interactions through a heterogeneous graph and uses word and style nodes as constrained auxiliary anchors rather than standalone lexical or stylistic shortcuts. Experiments on MCFEND show that SHIELD remains competitive with recent graph-enhanced detectors in the mixed-source setting. On the two controlled unseen-source-group subsets, SHIELD achieves the highest Avg.F1 among the evaluated baselines, and its F1 advantage persists after reducing target-label skew and changing the unseen target group. The results also show that general-purpose LLMs do not consistently perform well in this setting. Overall, the findings support the use of claim-reaction interactions, together with training-only semantic and stylistic anchors, for source-group generalization within the evaluated MCFEND source groups.

Although the controlled evaluations reduce label skew and add G1 as a second unseen target, they remain limited to two relatively small target groups from MCFEND and do not establish unrestricted generalization to arbitrary sources. The SnowNLP ablation supports the downstream utility of the sentiment feature, but we have not directly calibrated its scores against human annotations from translated and informal reply domains. In addition, the threshold-sensitivity analysis supports $k > 5$ as an empirical setting for interaction-based graph modeling, but the current evaluation still excludes documents with five or fewer valid replies. SHIELD therefore relies on a minimum amount of user feedback, which limits its use in cold-start and sparse-interaction scenarios.

The LLM-assisted reply selection is designed to be stance-agnostic and information-oriented, but the current study does not include a separate human-annotated audit of stance diversity, sentiment distribution, or separability changes after selection. Future work will explore domain-specific sentiment calibration, multi-modal extension using visual cues, human-audited reply-selection diagnostics, and few-shot adaptation using LLM-based reasoning when social reactions are sparse.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 72293575.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Rongfa Chen and Liping Chen; methodology, Rongfa Chen and Liping Chen; software, Rongfa Chen; validation, Rongfa Chen; formal analysis, Rongfa Chen; investigation, Rongfa Chen; resources, Daniel Zeng; data curation, Rongfa Chen and Xiuzhe Meng; writing—original draft preparation, Rongfa Chen; writing—review and editing, Rongfa Chen and Liping Chen; visualization, Rongfa Chen and Xiuzhe Meng; supervision, Daniel Zeng; project administration, Daniel Zeng; funding acquisition, Daniel Zeng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The MCFEND dataset analyzed during the current study is publicly available at https://github.com/TrustworthyComp/mcfend. Other datasets and codes generated during the current study are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable. This study does not involve human or animal subjects.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Reply Selection and LLM Evaluation

Appendix A.1 LLM-Assisted Reply Selection

For reply selection, we first remove invalid replies whose length is no more than five characters. If a news item has no more than 10 valid replies, all replies are retained. If it has more than 10 valid replies, an LLM selects at most 10 representative replies according to the prompt summarized below. This design is used to improve informational validity rather than to choose a particular stance. The prompt asks for replies that provide factual judgment, stance expression, logical questioning, or content-relevant evidence, and it does not ask the LLM to favor replies that support or oppose the claim.

Input: News title/content x and candidate replies $\{r_1, ..., r_n\}$
Task: Select at most 10 representative replies from the candidate list
Criteria: Relevance to the claim; semantic informativeness; coverage of informative viewpoints when present; usefulness for judging the claim-reaction relationship
Output: Indices of selected replies in the candidate list

The selected replies are identified by their indices in the candidate list. Invalid indices are discarded; if fewer than 10 valid replies are returned, the remaining slots are filled by randomly sampling from the unselected valid replies. If the LLM output is malformed or the request fails, random sampling is used as a fallback.

This reply-selection step does not use ground-truth labels, validation data, or target-group statistics. It is applied only to reduce redundant or irrelevant social reactions before graph construction. The selection strategy is stance-agnostic: replies may be retained whether they support, question, oppose, or neutrally discuss the claim, as long as they provide informative feedback. To reduce over-filtering risk, LLM-based selection is used only for news items with more than 10 valid replies; otherwise, all valid replies are kept. After simple normalization, including URL removal, whitespace and punctuation removal, and lowercasing, approximately 40% of news items contain at least one duplicate or near-duplicate reply. The median number of replies per news item is 12; therefore, selecting at most 10 informative replies retains most of the typical feedback scale while controlling graph size and computational cost. After reply selection, we further retain only news items with more than five valid replies, yielding the final 10,303 refined instances used in the

experiments. The selected subset should be interpreted as a compact, information-oriented approximation of the observable feedback set rather than a complete record of all user reactions.

Appendix A.2 Zero-Shot LLM Evaluation

Each LLM baseline is evaluated under the same zero-shot protocol. The input consists of the news text and, when available, its associated social reactions, provided within the supported context window of each LLM. The temperature is fixed to 0.1, and neither self-consistency voting nor majority voting is used. The prompt asks the model to classify the news as *Fake* or *Real* by considering source reliability, factual objectivity, emotional language, logical consistency, and the relationship between the claim and user reactions. The prompt is summarized in pseudo-code format as follows:

Input: News text x and optional social reactions r
Task: Judge whether x is *Fake* or *Real*
Criteria: Source reliability; factual objectivity; emotional language; logical consistency; claim-reaction relationship
Output:
```
Prediction: Fake/Real
Reason: brief explanation
```

The final label is extracted from the `Prediction` field. If this field is absent, the raw response is searched for explicit *Fake* or *Real* labels. Responses containing both labels, no recognizable label, or malformed content are marked as `unknown` and counted as incorrect predictions. Failed or timed-out requests are retried; unresolved cases are also treated as `unknown`. No training or validation samples from the unseen target source group are used for prompt design, model selection, or calibration.

References

1. Li Y, He H, Bai J, Wen D. MCFEND: a multi-source benchmark dataset for Chinese fake news detection. arXiv:2403.09092. 2024.
2. Zhou X, Zafarani R. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv. 2020;53(5):1–40. doi:10.1145/3395046.
3. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data. 2020;8(3):171–88.
4. Wang WY, Chang YC, Peng WC. Style-News: incorporating stylized news generation and adversarial verification for neural fake news detection. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); 2024 Mar 17–22; St. Julian's, Malta. p. 1531–41. doi:10.18653/v1/2024.eacl-long.92.
5. Madden R. A style-based approach for detecting COVID-19 fake news [master's thesis]. Dublin, Ireland: Technological University Dublin; 2023.
6. Lebernegg N, Eberl JM, Tolochko P. Do you speak disinformation? Computational detection of deceptive news-like content using linguistic and stylistic features. Digit J. 2025;13(8):1373–98. doi:10.1080/21670811.2024.2305792.
7. Salminen J, Mustak M, Jung SG, Makkonen H. Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models. J Market Anal. 2025;29(8):3846. doi:10.1057/s41270-025-00393-8.
8. Devlin J, Chang M-W, Lee K, Toutanova K. 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; 2019 Jun 2–9. Minneapolis, MN, USA. p. 4171–86.
9. Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: what we know about how BERT works. Trans Assoc Comput Linguist. 2020;8:842–66.

10. Nasir A, Wasim M, Nasir S. Credify: contextualized retrieval of evidence for open-domain fact verification. Knowl Inf Syst. 2025;67(7):5699–729. doi:10.1007/s10115-025-02400-x.
11. Aperstein Y, Gottlib A, Benita G, Apartsin A. Explainable semantic text relations: a question-answering framework for comparing document content. Information. 2025;16(12):1090.
12. Zhang W, Sheng Q, Alhazmi A, Li C. Adversarial attacks on deep-learning models in natural language processing: a survey. ACM Trans Intell Syst Technol. 2020;11(3):1–41.
13. Silva A, Luo L, Karunasekera S, Leckie C. Embracing domain differences in fake news: cross-domain fake news detection using multi-modal data. Proc AAAI Conf Artif Intell. 2021;35(1):557–65. doi:10.1609/aaai.v35i1.16134.
14. Mishra S, Shukla P, Agarwal R. Analyzing machine learning enabled fake news detection techniques for diversified datasets. Wirel Commun Mob Comput. 2022;2022(1):1575365. doi:10.1155/2022/1575365.
15. Wu L, Chen Y, Shen K, Guo X, Gao H, Li S, et al. Graph neural networks for natural language processing: a survey. Found Trends Mach Learn. 2023;16(2):119–328.
16. Monti F, Frasca F, Eynard D, Mannion D, Bronstein MM. Fake news detection on social media using geometric deep learning. arXiv:1902.06673. 2019.
17. Truica CO, Apostol ES, Marogel M, Paschke A. GETAE: graph information enhanced deep neural network ensemble ArchitecturE for fake news detection. Expert Syst Appl. 2025;275:126984.
18. Lee B, Cao D, Zhang T. MGMP: multi-granularity semantic relation learning and meta-path structure interaction learning for fake news detection. Appl Intell. 2025;55(7):655.
19. Xie K, Wang S. A survey on false information detection: from a perspective of propagation on social networks. arXiv:2506.18052. 2025.
20. Cui J, Kim K, Na SH, Shin S. Meta-path-based fake news detection leveraging multi-level social context information. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM); 2022 Oct 17–21; Atlanta, GA, USA.
21. Lakzaei B, Haghir Chehreghani M. Disinformation detection using graph neural networks: a survey. Artif Intell Rev. 2024;57(3):52.
22. Xie B, Ma X, Xue S, Yang J, Wu J, Fan H. Contrastive multi-knowledge graph learning for fake news detection. IEEE Trans Netw Sci Eng. 2025;12(5):3948–61. doi:10.1109/tnse.2025.3567296.
23. Feng X, Luo J, Yang Y, El Baz D. Health misinformation detection: approaches, challenges and opportunities. Inq J Health Care Organ Provis Financ. 2025;62:00469580251384784.
24. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3504–14. doi:10.1109/TASLP.2021.3124365.