Deciphering Rumors: A Multi-Task Learning Approach with Intent-aware Hierarchical Contrastive Learning

Chang Yang, Peng Zhang*, Hui Gao, Jing Zhang
College of Intelligence and Computing, Tianjin University, Tianjin, China
{yangchang, pzhang, hui_gao, zhang_jing}@tju.edu.cn

Abstract

Social networks are full of noise and misleading information, which poses a pressing and complex challenge for rumor detection. In this paper, we propose the Intent-Aware Rumor Detection Network (IRDNet), designed to address the challenges of subjectivity, robustness, and consistency in existing models. IRDNet uses a multi-task learning framework that integrates rumor detection and latent intent mining, which can discern multi-level semantic features and potential user intentions. In IRDNet, the multi-level semantic extraction module extracts sequential and hierarchical features to produce robust semantic representations. The intent-aware hierarchical contrastive learning module introduces two complementary strategies, event-level and intent-level. Event-level contrastive learning uses high-quality data augmentation and adversarial perturbations to enhance the robustness and consistency of the model. Intent-level contrastive learning utilizes an intent encoder to capture subjective intent and optimize homogeneity within the same intent while ensuring heterogeneity between different intents, thereby clearly distinguishing critical features from irrelevant elements. Experimental results verify that the model significantly improves the effect of early rumor detection and effectively solves the essential problems of the existing rumor detection field.

1 Introduction

The contemporary era of social networks witnesses unprecedented information dissemination, which concurrently leads to an accelerated spread of rumors. The complexity of this social phenomenon presents significant challenges in rumor detection (Gao et al., 2022). Fundamentally, the belief in rumors is contingent upon two criteria: "able to believe" and "willing to believe". The former involves the diversity of information in social networks and the difficulty of distinguishing its authenticity(Nickerson, 1998). The latter "willing to believe" means that during the communication process, under the influence of subjectivity, individuals rely on previous knowledge and experience as cognitive anchors to evaluate the authenticity of new information. This phenomenon, commonly called the anchoring effect(Tversky and Kahneman, 1974), assumes particular prominence in the propagation of rumors, as content aligned with individuals' existing cognition augments its dissemination. Therefore, rumor detection is not only a check on content authenticity but also a complex task involving human cognitive alignment. Similar to the idea of anchoring effect, contrastive learning (CL) emphasizes using known data as "anchors" in the process of information evaluation and generalizing it to new and unfamiliar data, which can be used to model the process of cognitive subjectivity. However, existing methods rely on the assumption that pairs of augmented data from the same sentence are semantically similar. As shown in the figure

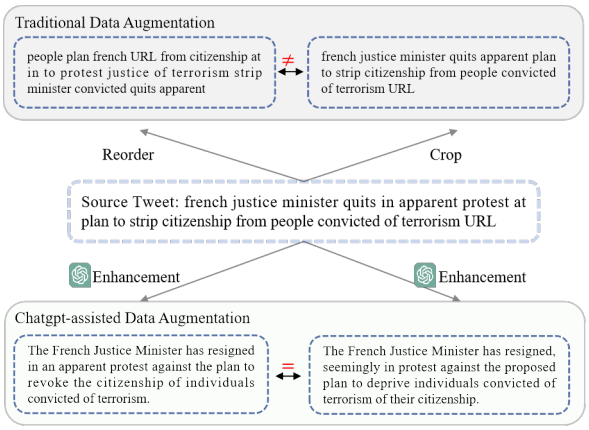

Figure 1: The results of traditional data augmentation methods are similar in semantic structures but inconsistent with the intentions of the information disseminator, while data augmentation based on large language models has been proven to be effective.

* Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4471–4483
November 12-16, 2024

1, this may cause contrastive learning to introduce potential noise or data that contradicts the intention of the information disseminator, ultimately leading to poor performance(Li et al., 2023). Traditional rumor detection methods often ignore the true intentions of information disseminators, rarely consider the importance of personal subjectivity, and have difficulty effectively capturing and aligning these implicit intentions. Moreover, distinguishing key features from irrelevant noise in text representation in noisy environments is a considerable challenge. Many methods have limited adaptability to the noisy social network environment, making it difficult to deal with the diversity of information effectively.

Inspired by this, we propose the Intent-Aware Rumor Detection Network (IRDNet), which designs a new contrastive learning method targeting human subjectivity, model robustness, and consistency of key features, and constructs cognitive anchors to mine the potential intentions of information disseminators. It includes the following key components:

In *Semantic Feature Extraction Module*, we design a semantic extractor that employs pre-trained models to enhance the model's semantic and contextual relevance handling. Additionally, it combines BiLSTM and capsule networks to achieve sequential and hierarchical feature learning, strengthening the model's ability to capture deep semantic features.

In *Intent-aware Hierarchical Contrastive Learning Module*, cognitive anchors are mainly implemented through two levels of contrastive learning: event-level contrastive learning and intent-level contrastive learning. In event-level contrastive learning, we combine adversarial training techniques to build high-quality data augmented representations to enhance the robustness of the model. In intent-level contrastive learning, by building intent-aware pairs and leveraging intent-level contrastive learning, it can separate individual intents and capture the main intent in a fine-grained manner while ensuring intent consistency.

In general, the main contributions of this paper are as follows:

- Aiming at the complex aspects of personal subjectivity, robustness and consistency in rumor detection, we design a multi-task learning framework that combines rumor detection and latent intent mining as key tasks by jointly optimizing self-supervised loss and supervised loss.

- We propose intent-aware hierarchical contrastive learning, which includes two complementary strategies: event-level and intent-level contrastive learning. It aims to model cognitive subjectivity through contrastive learning, enhance the robustness and consistency of model representation, and address existing challenges in rumor detection.

- We evaluate the proposed method using two real-world datasets and compare it with baseline methods. Experimental results verify the effectiveness of our method and demonstrate its significant advantages in early rumor detection tasks.

2 Related Work

Previous rumor detection methods primarily relied on feature engineering techniques to extract features from existing rumor instances (Horne and Adali, 2017; Castillo et al., 2011; Yang et al., 2012; Potthast et al., 2017; Wang, 2017). These studies focused on utilizing features such as post content, user profiles, and propagation patterns to train supervised classifiers. However, the effectiveness of this approach heavily depended on the quality of feature engineering, leading to potential issues with model generalization across diverse datasets.

Deep learning focuses on extracting deep key features, such as using recurrent neural networks (Ma et al., 2016; Asghar et al., 2021) and convolutional neural networks (Ajao et al., 2018; Liu and Wu, 2018; Yu et al., 2017) to extract important information from text content. In addition, capsule networks have been used to address the limitations of CNN (Zhao et al., 2018; Ren and Lu, 2022; Yang et al., 2023). Capsule networks use a dynamic routing mechanism to replace the maximum pooling operation of CNN to retain spatial information and capture key patterns in text. (Sabour et al., 2017; Mazzia et al., 2021). In recent advancements, graph neural network models have been employed to exploit valuable features from content semantics and propagation structures. These models encode conversation threads by modeling propagation trees (Ma et al., 2018b; Wu et al., 2015; Yang et al., 2024) and propagation graphs (Wei et al., 2022; Bian et al., 2020; Lin et al., 2021), resulting in higher-level representations. Nevertheless, when

data is limited in the early stages of rumor propagation, the above methods may not fully exploit their advantages(Hedderich et al., 2020).

Contrastive learning (CL) can alleviate the problem of data scarcity, especially by improving data quality and utilizing limited labeled data, and has gradually been applied to rumor detection(Lin et al., 2022; Xu et al., 2023; Gao et al., 2023; Cui and Jia, 2024). In addition, multi-task learning (MTL) is a strategy that improves the generalization ability of the model by leveraging the shared knowledge between multiple interrelated tasks(Caruana, 1997), which has also been widely explored and used in the field of rumor detection(Zhang et al., 2021; Zhang and Gao, 2024). Early studies(Kochkina et al., 2018; Ma et al., 2018a) mainly emphasize shared features and improve the effectiveness of rumor detection by promoting feature interactions between different tasks. Recent studies(Khandelwal, 2021; Yang et al., 2022; Ma et al., 2024) simultaneously consider related tasks, such as rumor detection and stance detection, to understand the text more comprehensively. Some studies (Cui and Yang, 2022; Zhou et al., 2022; Li et al., 2024) also attempt to apply MTL in multimodal environments, fusing multiple information sources such as text, images, or audio, aiming to more effectively capture the multidimensional properties of rumor propagation, thereby improving the overall performance of rumor detection.

3 Problem Statement

In the task of rumor detection, we define a series of events in the datasets, $E = \{E_1, E_2, E_3, ..., E_m\}$. Each event E_i is linked to a source tweet s_i and all associated comment contents, denoted as $E_i = \{s_i, x_1, x_2, ..., x_n\}$, where s_i can also be represented as x_0, and n represents the number of relevant comments in the discussion thread. The task goal is to build a robust rumor detection classifier by fully mining and learning the critical feature representations of rumor-related events on social networks, which can be expressed as $f : E_i \rightarrow Y_i$, where Y_i belongs to the categories of non-rumor, false rumor, true rumor, and unverified.

4 Methodology

As shown in Figure 2, we propose an intent-aware rumor detection network (IRDNet) using a multi-task training framework, including supervised semantic feature extraction tasks and self-supervised intent-aware hierarchical contrastive learning tasks and jointly optimize them by fusing supervised and self-supervised loss functions. Each module's principles and specific details will be introduced below.

4.1 Semantic Feature Extraction Module

In this module, we use BERTweet (Nguyen et al., 2020) as a semantic extractor, a language model pre-trained on a large-scale tweet corpus to enhance semantic and contextual processing capabilities. We further incorporate BiLSTM and Capsule Networks into our model to facilitate multi-level semantic extraction, thereby capturing critical text-based information more comprehensively.

We input the text sequence s_i into the semantic extractor, producing the output $e_i \in \mathbb{R}^{s \times d}$, where s denotes the sequence length. Subsequently, BiLSTM conducts sequential feature extraction to generate hidden state representations.

$$H_i = [\overrightarrow{LSTM}(e_i), \overleftarrow{LSTM}(e_i)] \in \mathbb{R}^{s \times 2h} \quad (1)$$

where $2h$ signifies the hidden state dimension of BiLSTM. For hierarchical feature extraction, we use a 1D convolutional layer for n-gram feature extraction obtaining primary capsules, $u_i = Conv(H_i) \in \mathbb{R}^{(s-k+1) \times k\_num}$, k is the convolution kernel size and k_num is the number of convolution kernels. Normalize using the squeeze function:

$$u_i = Squash(u_i) = \frac{||u_i||}{1 + ||u_i||^2} \frac{u_i}{||u_i||} \quad (2)$$

Next, the total input v_k of the next layer capsule is the weighted sum of primary capsule's prediction vectors $\hat{u}_i \in \mathbb{R}^{c_{in} \times (s-k+1) \times c_{dim}}$, c_{in} denotes the number of higher-level capsules, and c_{dim} represents the dimension of the capsules.

$$v_k = \sum_k c_{k|i} \cdot \hat{u}_{k|i}, \ \hat{u}_i = W_i^t u_i \quad (3)$$

where the predicted vector $\hat{u}_i$ is obtained through a projection process, and the coupling coefficients $c_k|i$ are derived through an iterative dynamic routing process, by applying the softmax function to $b_{k|i}$. And the values of $b_{k|i}$ are updated by:

$$c_{k|i} = softmax(b_{k|i}) \quad (4)$$

$$b_{k|i} = b_{k|i} + \hat{u}_{k|i} \cdot v_k \quad (5)$$

After r rounds of dynamic routing, the capsule network obtains the final output $v_k \in \mathbb{R}^{c_{in} \times c_{\text{dim}}}$, which encapsulates multi-level semantic features.

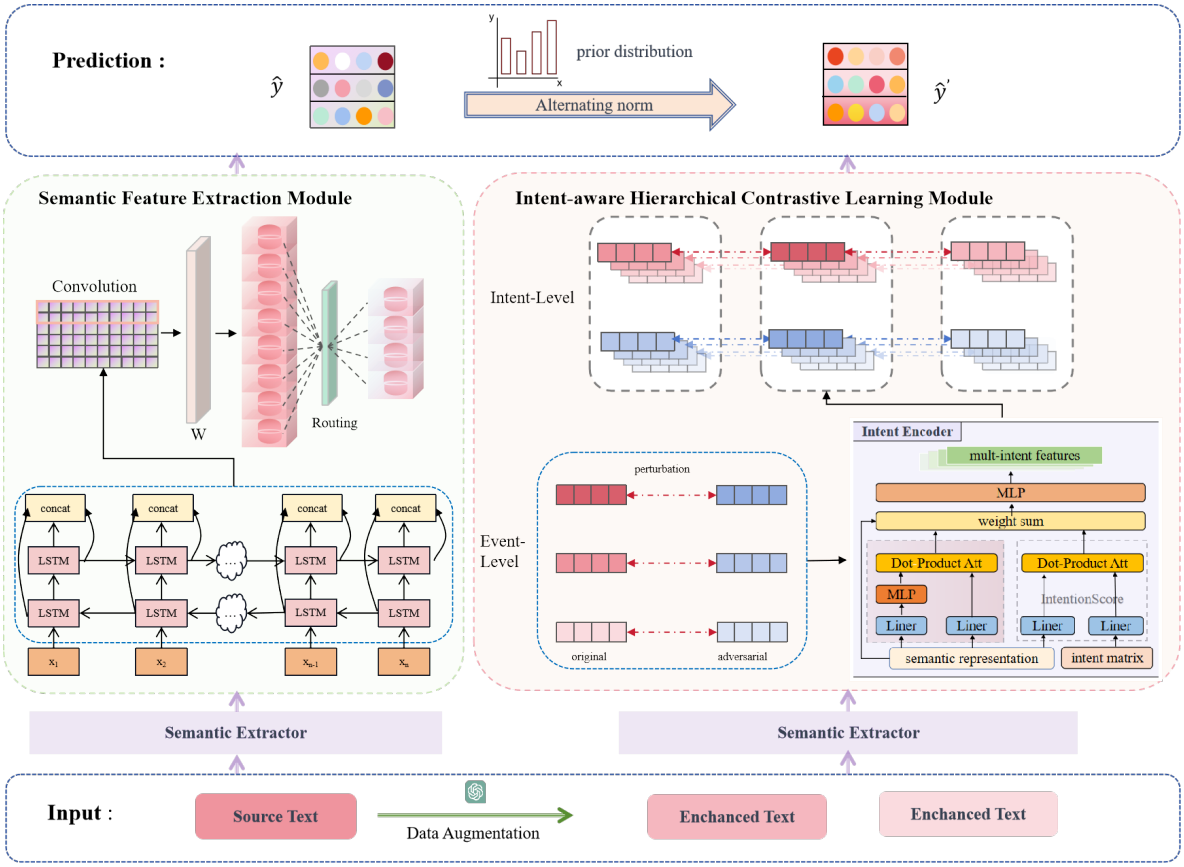

Figure 2: IRDNet, a multi-task learning framework that mainly includes supervised semantic feature extraction and self-supervised intent-aware hierarchical contrastive learning, captures key content features and potential intent features to enhance the model's robustness and consistency. In addition, post-processing model optimization techniques are used in our prediction stage, as shown in Appendix A

.

4.2 Intent-aware Hierarchical Contrastive Learning Module

In this section, we develop two complementary contrastive learning strategies: event-level contrastive learning and intent-level contrastive learning, aiming to enhance the model's ability to handle data noise in social networks and to understand the subjective intent of information disseminators. As shown in Figure 3, these strategies help the model obtain key contextual semantic information and further improve its ability to distinguish current noise.

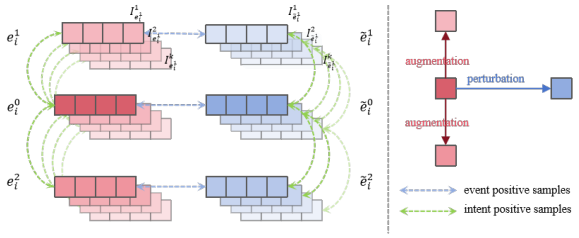

Figure 3: Details of two contrastive learning in intention-aware hierarchical contrastive learning

4.2.1 Event-Level Contrastive Learning

Data Augmentation: We utilize ChatGPT to enhance the original text data, aiming to eliminate irregularities and improve the overall quality and accuracy of the data. ChatGPT (Abdullah et al., 2022) employs Reinforcement Learning from Human Feedback (RLHF) to generate text data that is closely aligned with the user's context and intent, demonstrating excellent performance in natural language processing tasks (Feng et al., 2023). To effectively guide the generation process, we utilize task prompts designed by Yang et al. (2023). For each tweet's semantic representation e_i, where can also be expressed as e_i^0, two augmented sentences e_i^1, e_i^2 are generated using ChatGPT-based data augmentation.

Adversarial Training: We incorporate adversarial training techniques to enhance the model's ro-

bustness against attacks. The Fast Gradient Sign Method (Miyato et al., 2016) is utilized to generate adversarial examples by adding small perturbations based on the gradient of the loss function $\mathcal{L}(f_\theta(e_i + r), y_i)$.

$$r_i = \epsilon \cdot \text{sign}(\nabla \mathcal{L}(f_\theta(e_i), y_i)) \tag{6}$$

where r is the perturbation, ϵ represents the disturbance amplitude, f_θ represents the classifier with parameters θ and y_i is the label.

Event-Level Contrastive Learning: We construct training pairs to ensure that adversarial samples are close to original samples in vector space, so that the model can learn representations that are insensitive to noise, thereby improving robustness and generalization capabilities. Specifically, each sentence $e_i \in \{e_i^0, e_i^1, e_i^2\}$ generates its adversarial example $\tilde{e_i} \in \{\tilde{e_i^0}, \tilde{e_i^1}, \tilde{e_i^2}\}$. The positive pairs are formed as $\{(e_i, \tilde{e_i}) \mid i \in \text{batch}\}$. The negative pairs include $\{(e_i, \tilde{e_j}) \mid i \neq j, j \in \mathcal{N}(i)\}$, where $\mathcal{N}(i)$ represents the set of negative samples for e_i. Then the event-level contrastive loss $\mathcal{L}_{event}$ is formulated as follows:

$$\mathcal{L}_{event} = -\log \frac{\exp(\text{sim}(e_i, \tilde{e_i})/\tau)}{\sum\limits_{j \in \mathcal{N}(i)} \exp(\text{sim}(e_i, \tilde{e_j})/\tau)} \tag{7}$$

where τ is the temperature parameter. Since text inputs are typically discrete (one-hot vectors), we directly perturb the embedding matrix. The specific steps of the perturbation process are illustrated in the following pseudocode:

Algorithm 1 Event-Level Contrastive Learning

Input: Semantic representation $\{e_i^0, e_i^1, e_i^2\}_{i=1}^B$ with the label $\{y_i\}_{i=1}^B$, perturbation magnitude ϵ, negative samples $\mathcal{N}(i)$

Output: Event-level contrastive loss L_{event}

for $e_i \in \{e_i^0, e_i^1, e_i^2\}$ with label y_i **do**
    Forward Pass loss: $\mathcal{L}(f_\theta(e_i), y_i)$
    Gradient Computation: $\nabla_{e_i} \mathcal{L}(f_\theta(e_i), y_i)$
    Perturb: $r_i = \epsilon \cdot \text{sign}(\nabla_{e_i} \mathcal{L}(f_\theta(e_i), y_i))$
    Adversarial Generation: $\tilde{e_i} = e_i + r_i$
end for
Event-level contrastive loss using formula (7)
for e_i and adversarial example $\tilde{e_i}$ **do**
    Positive pairs: $(e_i, \tilde{e_i})$
    Negative pairs: $\{(e_i, \tilde{e_j}) \mid j \in \mathcal{N}(i))\}$
end for

4.2.2 Intent-Level Contrastive Learning

In this module, we design an intent encoder to obtain different intent representations by putting the original data and perturbation data into the intent encoder to capture the subjective intent of information disseminators in social networks. Through contrastive learning at the intent level, we can bring the positive samples closer and ensure that the augmented data and perturbation data maintain the consistency of the original intent.

Intent Matrix: To better comprehend the multiple meanings and latent motivations within the text, we introduce a parameter matrix $c_i = \{c_i^1, c_i^2, \ldots, c_i^K\}$, where K represents the number of latent intentions. This matrix can adaptively represent different latent intentions to deeply explore the multidimensional subjective value superposition process of information disseminators.

Saliency Weight: As a metric for assessing the importance of tweet content, it provides crucial clues for subsequent intention analysis and aids in reducing the impact of noise and irrelevant information. A non-linear transformation is introduced to enhance the model's expressive power, which is crucial for understanding and representing complex structures in the text. The calculation formula for Saliency Weight is as follows:

$$e_i^{key} = e_i + ReLU(W^{key} \cdot e_i) \tag{8}$$

$$e_i^{query} = W^{query} \cdot e_i \tag{9}$$

$$P_A(e_i) = Softmax\left(\frac{e_i^{key} \cdot (e_i^{query})^T}{\sqrt{d}}\right) \tag{10}$$

where e_i represents the embedding representation of text sequence x_i through the semantic extractor.

Intent Score: This metric measures the relevance of a tweet to different latent intentions. Specifically, for each intention, the score reflects the extent of association between the tweet and the intention. A low association with all intentions might be considered noise or an intention not yet learned by the model. To determine the likelihood of a text e_i being related to the k-th latent intention:

$$P_I^k(e_i) = \frac{\exp\left(\frac{W_e \cdot e_i \cdot c_i^{(k)^T}}{\sqrt{d}}\right)}{\sum_{k'=1}^{K} \exp\left(\frac{W_e \cdot e_i \cdot c_i^{(k')^T}}{\sqrt{d}}\right)} \tag{11}$$

where $P_I^k(e_i)$ is the probability of text e_i being related to the k-th latent intention, $c_i^{(k)}$ represents the

k-th latent intention vector, $\sqrt{d}$ is a scaling factor, and K is the total number of latent intentions.

Finally, the subjective intention representation of individuals in social networks is captured by leveraging intention interactions as follows:

$$I_{e_i}^k = ReLU\left(P_I^k(e_i) \cdot P_A(e_i) \cdot e_i + b_1\right) \quad (12)$$

Intent-Level Contrastive Learning: We obtain $6k$ intention feature representations through the intent encoder, resulting in $4k$ pairs of intent-aware positive samples, and negative samples are obtained by sampling $j \in \mathcal{N}(i)$. The intent-level contrastive loss function is as follows:

$$\mathcal{L}_{intent} = -\sum_{k=1}^{K}\left[\log \frac{\sum_{(x,y)\in A^k} \exp(f(x,y)/\tau)}{\sum_{j\in N'(i)} \exp(f(I_{e_i}^k, I_{e_j}^k)/\tau)}\right] \quad (13)$$

A^k represents the set of k-th intent positive sample pairs $\{(I_{e_i}, I_{e_i^1}), (I_{e_i}, I_{e_i^2}), (I_{\tilde{e}_i}, I_{\tilde{e}_i^1}), (I_{\tilde{e}_i}, I_{\tilde{e}_i^2})\}$. The final loss function includes the supervised loss function and the self-supervised contrastive learning loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda(\mathcal{L}_{event} + \mathcal{L}_{intent}) \quad (14)$$

Where $\mathcal{L}_{CE}$ is the loss function for supervised learning, λ is a hyperparameter that controls the contribution of the intent-aware hierarchical contrastive loss.

5 Experiments

5.1 Dataset and Experimental Setup

In this study, we use two public datasets in the field of rumor detection, Twitter15 and Twitter16 (Ma et al., 2017), covering four categories: true rumors (T), false rumors (F), unverified rumors (U), and non-rumors (N).

We conducted experiments on NVIDIA Tesla V100 GPU using the PyTorch framework with the following configuration: batch size set to 64 and an initial learning rate of 1e-5. To control the complexity of the model, we introduce an L2 regularization term with a weight attenuation coefficient of 0.001. To avoid overfitting, we implemented an early stopping strategy (Yao et al., 2007). We divided the dataset throughout the experiment into 70% training, 20% validation, and 10% testing. The detailed model parameter settings and hyperparameter sensitivity analysis are in Appendix B.

5.2 Baselines

We compare IRDNet with different baseline models, including content-based, structure-based, and contrastive learning-based methods:

(1) **GRU-RNN** (Ma et al., 2016): This model employs a Recurrent Neural Network (RNN) for event classification, distinguishing between rumors and non-rumors.

(2) **dEFEND** (Shu et al., 2019): The dEFEND model employs attention mechanisms to capture semantic correlations between the source tweet content and user comments for rumor detection.

(3) **RvNN** (Ma et al., 2018b): The RvNN model captures the structure features within the propagation tree using a Recursive Neural Network for rumor detection.

(4) **BiGCN** (Bian et al., 2020): The BiGCN model captures the structural features of rumor propagation by constructing a bidirectional graph structure and leveraging a graph convolutional network (GCN).

(5) **TextGCN** (Yao et al., 2019): The TextGCN model utilizes graph convolutional network models to learn text representations for text classification tasks.

(6) **CICAN** (Yang et al., 2023): The CICAN model captures multiple semantic content features and different semantic structure features in text for rumor detection.

(7) **GACL** (Sun et al., 2022): The GACL model incorporates graph-based adversarial contrastive learning, utilizing rumor propagation structure information for rumor detection.

(8) **RAGCL** (Cui and Jia, 2024): The RAGCL model utilizes graph contrastive learning with adaptive view augmentation guided by node centralities.

5.3 Experimental Results

In this section, we discuss the comparative analysis of the performance of our proposed IRDNet model against other methods, as summarized in Table 1. The IRDNet model delivers significant performance improvements on the Twitter15 and Twitter16 datasets, achieving 91.7% and 90.9% accuracy, respectively.

Content-based models, such as GRU-RNN and dEFEND, employ deep learning methods to consider the relationship between reviews and the original text, such as the attention mechanism and the recurrent neural network (RNN). However, these models often overlook extracting deep semantic

Method	Module			Twitter15					Twitter16				
	Content -based	Structure -based	CL -based	ACC	F1				ACC	F1			
					N	F	T	U		N	F	T	U
GRU-RNN	✓	-	-	0.708	0.739	0.68	0.699	0.681	0.683	0.622	0.671	0.697	0.664
dEFEND	✓	-	-	0.731	0.631	0.646	0.617	0.668	0.721	0.649	0.603	0.611	0.637
RvNN	✓	✓	-	0.713	0.702	0.691	0.746	0.654	0.737	0.662	0.743	0.801	0.768
BiGCN	✓	✓	-	0.798	0.716	0.758	0.843	0.876	0.803	0.792	0.788	0.796	0.814
TextGCN	✓	✓	-	0.83	0.801	0.859	0.842	0.794	0.846	0.816	0.832	0.824	0.801
CICAN	✓	✓	-	0.855	0.816	0.811	0.870	0.901	0.840	0.838	0.826	0.802	0.862
GACL	✓	-	✓	0.861	0.872	0.869	0.832	0.851	0.88	0.873	0.884	0.901	0.867
RAGCL	✓	-	✓	0.867	0.891	0.867	0.869	0.835	0.905	0.836	**0.923**	**0.963**	**0.882**
IRDNet	✓	-	✓	**0.917**	**0.914**	**0.928**	**0.905**	**0.897**	**0.909**	**0.891**	0.906	0.903	0.871

Table 1: Rumor detection results on Twitter15 and Twitter16 datasets.

features, resulting in relatively subpar performance. In contrast, IRDNet incorporates a multi-level semantic extraction module, enabling it to capture global-local relationships effectively.

Structure-based models, TextGCN, CICAN, and BiGCN, outperform RvNN due to their superior capability to model the complex relationships among structure features. However, these structure-based models exhibit lower robustness and are more vulnerable to malicious user attacks, leading to misleading results and consequently impacting the accuracy of rumor detection.

CL-based models, GACL, RAGCL and IRDNet, exhibit significant performance improvement attributable to two pivotal factors. Firstly, contrastive learning enhances the model's robustness, making it more resistant to malicious attacks. Secondly, contrastive learning effectively highlights the commonalities within the same category and differences across distinct categories by providing rich self-supervised signals, thereby enriching the model training process. In contrast, the IRDNet model utilizes hierarchical contrastive learning, including event-level and intent-level strategies. Event-level contrastive learning improves robustness by using adversarial training to make the model resistant to noise. Intent-level contrastive learning can capture user intent and keep the detected rumors consistent even in noisy data.

5.4 Ablation Experiments

We conducted a series of ablation studies to evaluate the contribution of each component in the IRDNet model. The specific configurations are as follows:

(1) **W/o BERTweet**: The BERTweet component in the Semantic Feature Extraction Module is replaced by the standard BERT.

(2) **W/o Cap**: The Capsule Network component in the Semantic Feature Extraction Module is removed.

(3) **W/o BiLSTM**: The BiLSTM component in the Semantic Feature Extraction Module is removed.

(4) **W/o IHCLM**: The Intent-aware Hierarchical Contrastive Learning Module is removed from the model.

(5) **W/o ECL**: The Event-level Contrastive Learning component in the Intent-aware Hierarchical Contrastive Learning Module is removed.

(6) **W/o ICL**: The intent-level contrastive learning component in the Intent-aware Hierarchical Contrastive Learning Module is removed.

Models	Twitter15-acc	Twitter16-acc
IRDNet	0.917	0.909
W/o BERTweet	0.884$_{-0.033}$	0.870$_{-0.039}$
W/o BiLSTM	0.893$_{-0.024}$	0.900$_{-0.009}$
W/o Cap	0.889$_{-0.028}$	0.887$_{-0.022}$
W/o IHCLM	0.890$_{-0.027}$	0.877$_{-0.032}$
W/o ECL	0.897$_{-0.020}$	0.882$_{-0.027}$
W/o ICL	0.893$_{-0.024}$	0.884$_{-0.025}$

Table 2: Ablation experiment results

As indicated by the results from the ablation experiments in Table 2, each module significantly impacts the model's performance. Specifically, the W/o BERTweet configuration has a substantial effect, resulting in performance reductions of 3.3% and 3.9% on the two datasets, respectively. This highlights the advantage of BERTweet over traditional BERT in capturing text semantics and

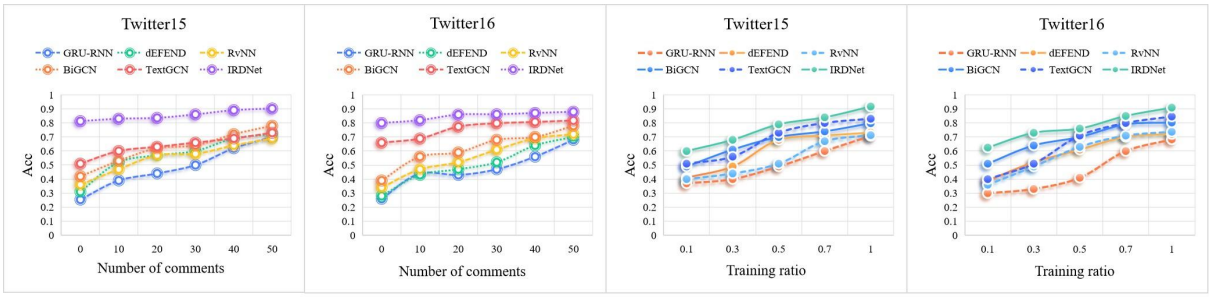

Figure 4: Early rumor detection experimental results with different number of comments (a, b) and different training ratios (c, d).

managing contextual relevance, while the Capsule network and BiLSTM contribute to multidimensional feature extraction, aggregation of input information from diverse perspectives, and capture of high-level semantic information. Removing the intention-level contrastive learning module (W/o IHCLM) resulted in a decrease in accuracy of 2.7% and 3.2%, highlighting the important role of this module in improving the model's adaptability to noisy data and its ability to capture semantic consistency. In contrastive learning strategies, removing event-level contrastive learning (W/o ECL) and intent-level contrastive learning (W/o ICL) both resulted in varying degrees of decreased accuracy.

It is proven that by utilizing a multi-level, fine-grained contrastive learning strategy, the model's robustness to noisy data is improved, and its ability to capture and understand the subjective intentions of information disseminators is also significantly enhanced.

5.5 Early Rumor Detection

Timely identification and mitigation are crucial for rumor detection. As shown in Figure 6, we utilize two methods to evaluate the model's early rumor detection performance. First, we evaluate the performance of the model when the number of comments gradually increases. The IRDNet model shows good accuracy even in the early stage with limited comments. This is attributed to the efficient extraction of semantic features and the advantages of event-level contrastive learning with adversarial training. Although the semantic noise increases as more information spreads in social networks, the IRDNet model always maintains excellent performance, highlighting its strong generalization ability.

Another method is to precisely change the ratio of the training dataset (0.1, 0.3, 0.5, 0.7), while fixing the validation ratio at 0.2 and the rest is allocated to the test set. Intent-level contrastive learning, by emphasizing the commonality between intent categories and the distinction between different intent categories, obtains more context-related semantic information from intent features, further improving our ability to distinguish current noise terms. Models based on graph neural networks (BiGCN and TextGCN) perform poorly, especially in scenarios with limited training data. When the training ratio is 0.1, our model shows a significant accuracy improvement of 10% to 30% compared with other models, demonstrating the superiority of our method.

6 Conclusion and Future Work

This paper proposes an intent-aware rumor detection network (IRDNet) that combines supervised semantic feature extraction with self-supervised intent-aware hierarchical contrastive learning in a multi-task framework to address the challenges of subjectivity, robustness, and consistency in the current field. It contains two modules: the semantic feature extraction module captures multi-level semantic information and enhances the model's ability to understand context. The intent-aware hierarchical contrastive learning module introduces two strategies: event-level contrastive learning enhances robustness through data augmentation and adversarial training, and intent-level contrastive learning captures subjective intent features by designing an intent encoder, ensuring intent consistency and improving feature discrimination while maintaining robustness. Experimental results show that IRDNet performs better noise resistance and critical feature capture than baseline methods in comparative experiments and early rumor detec-

tion.

Future research will focus on the following two aspects. First, the proposed contrastive learning approach will be extended to a broader range of classification tasks to verify its generality and effectiveness in diverse tasks, expanding its applicability. Second, efforts will focus on addressing the current limitations of large language models (LLMs) in rumor detection. Specifically, we will work towards enhancing the robustness and detection accuracy of LLMs in dynamic misinformation environments, while also ensuring higher reliability and consistency in rumor detection tasks.

7 Limitations

Despite the progress made with the Intent-Aware Rumor Detection Network (IRDNet), some limitations still exist:

Computational Complexity: The multi-task learning strategy integrating language model and contrastive learning increases the computational complexity and resource requirements of IRDNet. Although negative sampling and parameter freezing are employed to reduce complexity, real-time deployment and scalability in environments with limited computing resources may still present challenges.

Generality in Multilingual Environments: IRDNet's current training and evaluation focus primarily on English datasets. Its effectiveness and robustness in multilingual environments or non-English social media platforms have yet to be thoroughly evaluated. Future research will explore other languages and cross-language applications to improve the model's performance across diverse linguistic contexts.

8 Acknowledgements

This work is supported in part by the Natural Science Foundation of China (grant No. 62276188). We gratefully acknowledge their support.

9 Ethical Statement

This study aims to advance the understanding of rumor detection and promote the advancement of social media and information dissemination. We solemnly promise to abide by ethical principles throughout the research process, protect the rights of data participants, and ensure compliance with and transparency regarding data use.

References

Malak Abdullah, Alia Madain, and Yaser Jararweh. 2022. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. IEEE.

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2018. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society*, pages 226–230.

Muhammad Zubair Asghar, Ammara Habib, Anam Habib, Adil Khan, Rehman Ali, and Asad Khattak. 2021. Exploring deep neural networks for rumor detection. *Journal of Ambient Intelligence and Humanized Computing*, 12:4315–4333.

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 AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

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

Chaoqun Cui and Caiyan Jia. 2024. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 73–81.

Xinyu Cui and Li Yang. 2022. Fake news detection in social media based on multi-modal multi-task learning. *International Journal of Advanced Computer Science and Applications*, 13(7).

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957*.

Li Gao, Lingyun Song, Jie Liu, Bolin Chen, and Xuequn Shang. 2022. Topology imbalance and relation inauthenticity aware hierarchical graph attention networks for fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4687–4696.

Yuan Gao, Xiang Wang, Xiangnan He, Huamin Feng, and Yongdong Zhang. 2023. Rumor detection with self-supervised learning on texts and social graph. *Frontiers of Computer Science*, 17(4):174611.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Menglin Jia, Austin Reiter, Ser-Nam Lim, Yoav Artzi, and Claire Cardie. 2021. When in doubt: Improving classification performance with alternating normalization. *arXiv preprint arXiv:2109.13449*.

Anant Khandelwal. 2021. Fine-tune longformer for jointly predicting rumor stance and veracity. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 10–19.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*.

Jun Li, Yi Bin, Liang Peng, Yang Yang, Yangyang Li, Hao Jin, and Zi Huang. 2024. Focusing on relevant responses for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*.

Xuewei Li, Aitong Sun, Mankun Zhao, Jian Yu, Kun Zhu, Di Jin, Mei Yu, and Ruiguo Yu. 2023. Multi-intention oriented contrastive learning for sequential recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 411–419.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Guang Chen. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. *arXiv preprint arXiv:2204.08143*.

Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on twitter with claim-guided hierarchical graph attention networks. *arXiv preprint arXiv:2110.04522*.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Guanghui Ma, Chunming Hu, Ling Ge, and Hong Zhang. 2024. Dsmm: A dual stance-aware multi-task model for rumour veracity on social networks. *Information Processing & Management*, 61(1):103528.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018a. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*, pages 585–593.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018b. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Vittorio Mazzia, Francesco Salvetti, and Marcello Chiaberge. 2021. Efficient-capsnet: Capsule network with self-attention routing. *Scientific reports*, 11(1):14634.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Hao Ren and Hong Lu. 2022. Compositional coding capsule network with k-means routing for text classification. *Pattern Recognition Letters*, 160:1–8.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *Advances in neural information processing systems*, 30.

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

Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2789–2797.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Lingwei Wei, Dou Hu, Wei Zhou, and Songlin Hu. 2022. Uncertainty-aware propagation structure reconstruction for fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2759–2768.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE.

Yingrui Xu, Jingyuan Hu, Jingguo Ge, Yulei Wu, Tong Li, and Hui Li. 2023. Contrastive learning at the relation and event level for rumor detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chang Yang, Xia Yu, JiaYi Wu, BoZhen Zhang, and HaiBo Yang. 2024. Graph-aware multi-feature interacting network for explainable rumor detection on social network. *Expert Systems with Applications*, 249:123687.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7.

Peng Yang, Juncheng Leng, Guangzhen Zhao, Wenjun Li, and Haisheng Fang. 2023. Rumor detection driven by graph attention capsule network on dynamic propagation structures. *The Journal of Supercomputing*, 79(5):5201–5222.

Ruichao Yang, Jing Ma, Hongzhan Lin, and Wei Gao. 2022. A weakly supervised propagation model for rumor verification and stance detection with multiple instance learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1761–1772.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A convolutional approach for misinformation identification. In *IJCAI*, pages 3901–3907.

Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Multi-modal meta multi-task learning for social media rumor detection. *IEEE Transactions on Multimedia*, 24:1449–1459.

Xuan Zhang and Wei Gao. 2024. Predicting viral rumors and vulnerable users with graph-based neural multi-task learning for infodemic surveillance. *Information Processing & Management*, 61(1):103520.

Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. 2018. Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*.

Honghao Zhou, Tinghuai Ma, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2022. Mdmn: Multi-task and domain adaptation based multi-modal network for early rumor detection. *Expert Systems with Applications*, 195:116517.

A Model Optimization: Classification with Alternating Normalization

We employ the Classification with Alternating Normalization (CAN) (Jia et al., 2021) to enhance the prediction accuracy of low-confidence samples and ensure the average of all probability distributions aligns with the prior distribution (can be estimated directly from the training dataset). This method quantifies the ambiguity of the predicted class probability distribution, uses a threshold to distinguish high-confidence and low-confidence samples, and applies alternating normalization to rescale predictions for low-confidence instances, i.e., alternately normalizing across samples in each category and across categories in each sample. The measure of uncertainty in the prediction result, $H_{\text{top-k}}$, utilizes the top-k probability values to compute the entropy as follows:

$$H_{\text{top-k}}(p) = -\frac{1}{log\ k}\sum_{i=1}^{k}\hat{p}_i \log_p \hat{p}_i \qquad (15)$$

Where $\hat{p}_i$ is the normalized probability distribution of the first k probability values, $\hat{p}_i = p_i / \sum_{i=1}^{k} p_i$, and $1/log\ k$ is to limit the final indicator value to $[0, 1]$.

Assuming there are N samples, among which n are high-confidence with probabilities $p^{(1)}, ..., p^{(n)}$. The low-confidence samples are then corrected against the high-confidence samples by alternately normalizing instances of each category and categories of each instance. Specifically, for each low-confidence prediction $p^{(j)}, j \in (n+1, \ldots, N)$, we perform normalization across instances for each category together with the high-confidence prediction as follows:

$$p^{(k)} = \frac{p^{(k)} \odot \tilde{p}}{\bar{p}},\ \bar{p} = \frac{1}{n+1}(p^{(j)} + \sum_{i=1}^{n} p^{(i)}) \qquad (16)$$

Where $k \in (1, 2, \cdots, n, j)$, After the above operations, it is possible that the sum of $p^{(k)}$ may not equal to 1, so we normalize them as follows:

$$p^{(k)} \leftarrow \frac{p_i^{(k)}}{\sum_{i=1}^{m} p_i^{(k)}} \tag{17}$$

Notably, this correction process adjusts each sample independently, and the probability distributions of both the high-confidence and low-confidence samples are updated. However, we ultimately keep the updated probability distribution as the corrected result for the original low-confidence samples.

Next, we conducted an ablation experiment on CAN. As shown in 3, the experiment highlights the beneficial effect of decisions based on prior knowledge of model generalization capabilities.

Models	Twitter15-acc	Twitter16-acc
IRDNet	0.917	0.909
W/o CAN	0.913$_{-0.004}$	0.901$_{-0.008}$

Table 3: Ablation experiment results of CAN

B Model Experimental Configuration and Hyperparameter Sensitivity Analysis

B.0.1 Model Experimental Configuration

In the semantic feature extraction module, the hidden size of the BERTweet layer is set to 768 in bertweet-base and 1024 in bertweet-large, with weights frozen to prevent overfitting. We experimented with both bertweet-base and bertweet-large models[1] and found that bertweet-large performed better. The BiLSTM layer consists of 256 hidden units. The size of the convolutional kernels (k) is {2, 3, 4}. The number of convolutional kernels (k_num) is fixed at 256. The higher-level capsule layer contains 4 capsules (c_{in}). The output dimension of each capsule (c_{dim}) is set to 50. Dynamic routing is performed for 3 iterations ($r = 3$) to aggregate capsule outputs effectively.

In the intent-aware hierarchical contrastive learning module, we implement event-level and intent-level contrastive learning. For event-level contrastive learning, we use a perturbation magnitude $\epsilon = 0.001$ to generate adversarial examples, with negative samples $\mathcal{N}(i) = 5$ per positive pair to ensure robust learning. For intent-level contrastive learning, the intent encoder is configured with $K = 4$ latent intentions, each represented by a 64 dimensional vector and 10 negative samples per positive pair.

During training, we use the cross-entropy loss function to measure the discrepancy between predicted results and true labels, combined with the event-level and intent-level contrastive loss functions, weighted by a hyperparameter (λ). An early stopping strategy with a patience of 10 epochs prevents overfitting and ensures generalizability.

B.0.2 Hyperparameter Sensitivity Analysis

We present the impact of crucial hyperparameters on our experiments, including the number of intentions (k) and CL weight parameter (λ).

Number of Intentions (k)

The number of intents (k) is key to intent-level contrastive Learning. Figure 5 illustrates the impact of different numbers of intents on the two datasets, Twitter15 and Twitter16. When $k = 1$, the model's performance is subpar. As k increases, the model's performance improves significantly, reaching an optimal performance at $k = 4$. This suggests that employing multiple intentions better captures the semantics and information embedded in the text.

However, performance degrades as the number of intents increases (i.e., k > 4). This degradation can be attributed to the intent decomposition being too fine-grained, and the information fragmentation is severe, which limits the model's expressive power.

Interestingly, there is a trend of slightly improved performance in $k = 20$. This suggests that additional intents may still capture some useful semantic variation, albeit with diminishing returns. However, given the increased computational complexity and potential overfitting risk associated with higher k values, we set $k = 4$ as the optimal number of intents.

CL Weight Parameter (λ)

Within the loss function, the weight parameter (λ) that controls the contrastive learning (CL) loss is another key hyperparameter. We note that $\lambda = 0.05$ yields the optimal performance on the Twitter16 dataset, and $\lambda = 0.01$ is optimal for the Twitter15 dataset. However, the model's performance deteriorates when λ exceeds its optimal value. This phenomenon arises because an elevated weight on the contrastive loss can compromise the

[1] https://huggingface.co/docs/transformers/model_doc/bertweet

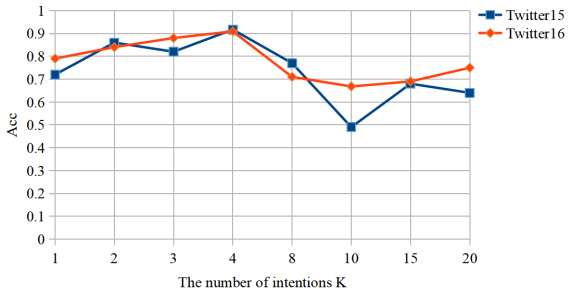

Figure 5: Sensitivity analysis of the number of intentions

quality of training for the classification task, resulting in decreased performance.

C Visualization Experiments

In the visualization experiment of event-level adversarial training, we use the confusion matrix to highlight the frequency of label transitions among key categories (non-rumor, true, false, unverified) before and after perturbation to compare the distribution of original labels with those resulting from perturbations.

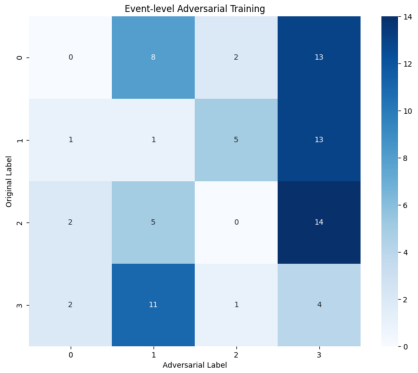

Figure 6: The impact of introducing perturbed samples in event-level contrastive learning on model robustness.

The heatmap indicates that the categories are frequently misclassified as "unverified (3)" after minor perturbations. This phenomenon can be attributed to the model's sensitivity to minor perturbations, where the adversarially generated samples may attenuate the critical features that originally distinguished these categories. Consequently, the model fails to effectively capture the core semantic features of the original categories, making it more likely to classify ambiguous information as belonging to an uncertain category, indicating that the model demonstrates insufficient robustness and contextual comprehension when confronted with semantic ambiguity. This behavior also aligns with human cognitive tendencies where slight shifts in wording or tone can provoke doubt, especially in the absence of clear evidence or authoritative support. Even if the information itself is accurate, if its description becomes ambiguous, individuals may lean towards considering it as unverified.