

SARC: Sentiment-Augmented Deep Role Clustering for Fake News Detection

Jingqing Wang
College of Computer Science,
Chongqing University
Chongqing, China
wangjq972@gmail.com

Jiaxing Shang*
Department of Computer Science,
University of Exeter
Exeter, UK
j.shang@exeter.ac.uk

Rong Xu
College of Computer Science,
Chongqing University
Chongqing, China
rongxu126@gmail.com

Fei Hao
School of Artificial Intelligence and
Computer Science, Shaanxi Normal
University
Xi'an, China
feehao@gmail.com

Tianjin Huang
Department of Computer Science,
University of Exeter
Exeter, UK
t.huang2@exeter.ac.uk

Geyong Min
Department of Computer Science,
University of Exeter
Exeter, UK
g.min@exeter.ac.uk

Abstract

Fake news detection has been a long-standing research focus in social networks. Recent studies suggest that incorporating sentiment information from both news content and user comments can enhance detection performance. However, existing approaches typically treat sentiment features as auxiliary signals, overlooking role differentiation, that is, the same sentiment polarity may originate from users with distinct roles, thereby limiting their ability to capture nuanced patterns for effective detection. To address this issue, we propose **SARC**, a **Sentiment-Augmented Role Clustering** framework which utilizes sentiment-enhanced deep clustering to identify user roles for improved fake news detection. The framework first generates user features through joint comment text representation (with BiGRU and Attention mechanism) and sentiment encoding. It then constructs a differentiable deep clustering module to automatically categorize user roles. Finally, unlike existing approaches which take fake news label as the unique supervision signal, we propose a joint optimization objective integrating role clustering and fake news detection to further improve the model performance. Experimental results on two benchmark datasets, RumourEval-19 and Weibo-comp, demonstrate that SARC achieves superior performance across all metrics compared to baseline models. The code is available at: <https://github.com/jxshang/SARC>.

CCS Concepts

• **Information systems** → **Data mining**; **Social networks**; • **Computing methodologies** → **Neural networks**.

Keywords

Fake News Detection; User Sentiment; Deep Clustering; Social Networks

ACM Reference Format:

Jingqing Wang, Jiaxing Shang, Rong Xu, Fei Hao, Tianjin Huang, and Geyong Min. 2026. SARC: Sentiment-Augmented Deep Role Clustering for Fake News Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (WSDM'26)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

The rapid advancement of the Internet and mobile communication technologies has fundamentally reshaped human communication patterns, positioning social media platforms as the dominant channels for information exchange [1–3]. By removing the temporal and spatial constraints of information dissemination, these platforms enable instant communication across global audiences, fostering positive societal impacts such as advancing social marketing [4, 5] and promoting social equality [6]. However, the same mechanisms that facilitate rapid and wide-reaching communication also create fertile ground for the proliferation of fake news, which can undermine public trust and distort social discourse [7, 8]. With the emergence of large language models (LLMs), the challenge has intensified, as fake news content has become increasingly sophisticated and harder to detect [9, 10]. Consequently, fake news detection on social networks has received growing attention in recent years. Addressing this challenge demands effective detection methods capable of operating in the dynamic and complex environment of online interactions. Most state-of-the-art approaches are based on deep learning technologies and can be broadly categorized into content-based, social-context-based, and LLM-based methods. For instance, Hu et al. [11] constructed a topic-entity-sentence heterogeneous graph enhanced with knowledge graphs to reveal differences between true and fake news. Cui et al. [12] proposed RAGCL, an adaptive graph contrastive learning model that captures structural features of rumor propagation trees through adaptive augmentation and contrastive learning. Zhou et al. [13] integrated LLM-based retrieval of up-to-date knowledge with small language models for collaborative fake news detection.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WSDM'26, Boise, Idaho, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

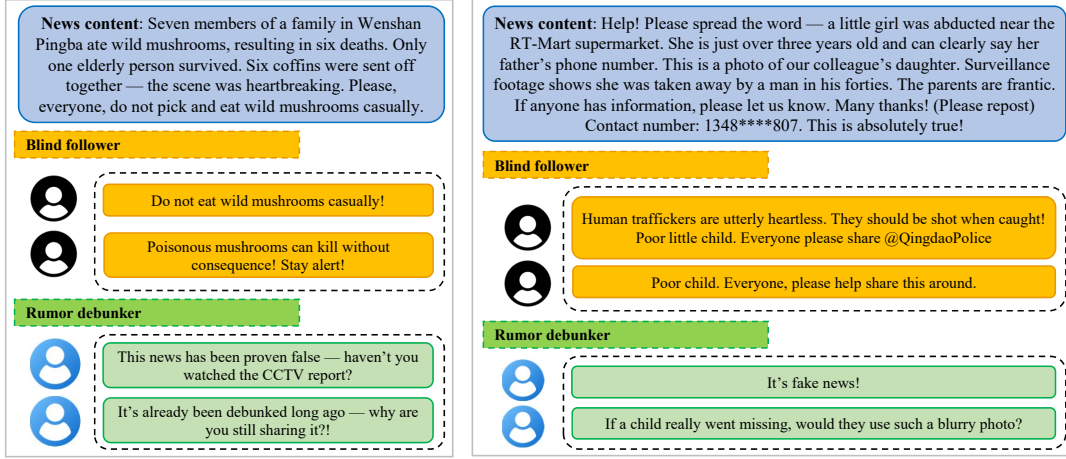


Figure 1: Motivational example showing two pieces of fake news and user comments from the Weibo-comp dataset.

Motivation. Despite notable progress achieved by deep learning-based fake news detection models, most existing studies still focus primarily on the semantic features of news content or shallow interaction signals between news content and user comments. Recent research has demonstrated that incorporating sentiment information from news content and user comments can provide valuable clues for identifying news authenticity [14–16]. For instance, Zhang et al. [17] extracted sentiment features representing dual emotions between news content and user comments to improve fake news detection. Jia et al. [18] jointly modeled sentiment and stance features, proposing a multi-task learning framework that enhances detection performance by training additional sentiment and stance classifiers. Zhang et al. [19] introduced a graph attention network-based model to capture mutual influences of sentiment between sentences within long-form news for fake news detection. However, these sentiment-based approaches typically treat sentiment and stance information as auxiliary features, overlooking the critical fact of *role differentiation*, that is, the same sentiment polarity may originate from users with distinct roles. As a motivational example, Figure 1 illustrates two fake news cases and their associated user comments from the Weibo-comp dataset, where comments expressing similar sentiment polarity (negative) stem from different user roles. Specifically, “blind followers” express negative emotions (e.g., anger) toward the news content without questioning its authenticity, whereas “rumor debunkers”, despite sharing a similar negative sentiment, explicitly express doubt about the news’s veracity. Neglecting such nuanced patterns can substantially limit the ability of existing models to detect fake news effectively.

Challenge. To incorporate the roles behind user comments into fake news detection, two critical challenges must be addressed. *Challenge 1: how to identify user roles when ground-truth annotations are unavailable?* To the best of our knowledge, none of the existing fake news datasets provide role labels for user comments. While stance detection methods [18, 20] implicitly categorize user comments into predefined roles (e.g., supporter or opponent), such roles are hard-coded and cannot be flexibly adapted to specific tasks

or datasets. *Challenge 2: how to effectively integrate the role identification task into the fake news detection pipeline to improve overall performance?* The key to addressing this challenge lies in designing an effective framework that jointly optimizes the role identification and fake news detection tasks.

Solution. To address the above challenges, we propose **SARC**, a **Sentiment-Augmented deep Role Clustering** framework that automatically identifies user roles in news comment sections to enhance fake news detection. Specifically, SARC incorporates an unsupervised, learnable deep role clustering module that jointly leverages sentiment features and textual content of comments to model collective behavioral patterns of users. By adopting an unsupervised deep clustering approach, SARC overcomes the limitations of traditional supervised role classification methods, which heavily rely on annotated datasets and suffer from label scarcity. Moreover, SARC introduces an integrated loss function that jointly optimizes the clustering loss and the fake news detection loss, thereby tightly coupling role identification with fake news detection and improving both accuracy and generalization. The SARC framework consists of four core components: (1) Initial Feature Representation Module, which establishes semantic foundations for news text and user comments using pretrained word embeddings while deriving sentiment polarity features for individual users; (2) Text Encoding Module, which utilizes dual-channel BiGRU networks with self-attention to extract deep semantic features from news content and temporal patterns from comments; (3) Dynamic Role Clustering Module, which creatively integrates differentiable deep clustering algorithms to automatically partition sentiment-augmented comment features into latent user behavioral patterns, while adaptively optimizing cluster distributions; (4) News Classification Module, which fuses news semantics with role features via feature concatenation and applies multilayer perceptrons for final classification.

The main contributions of this paper are as follows:

- We provide a novel perspective by modeling collective cognition in news comment sections through user roles. Unlike prior works that treat comments merely as auxiliary textual features, SARC conceptualizes role differentiation as

dynamic representations of group cognition, quantifying user affiliations (i.e., roles) via learnable role prototype vectors (i.e., cluster centers).

- We propose a differentiable deep clustering algorithm for the unsupervised categorization of user roles. In contrast to existing methods that rely solely on fake news labels for supervision, SARC introduces a joint optimization objective that integrates role clustering with fake news detection. By jointly optimizing clustering and detection losses, SARC tightly couples role identification with detection, thereby enhancing both accuracy and generalization.
- We conduct comprehensive experiments on two real-world datasets, comparing SARC with state-of-the-art methods. Results demonstrate that SARC achieves substantial improvements, ranging from 4.9% to 15.3% in detection accuracy over competitive baselines. Ablation studies further confirm the effectiveness of the proposed modules.

2 Related Work

2.1 Fake News Detection

In recent years, fake news detection on social media has garnered significant attention. Based on the types of information utilized by detection models, existing research can be broadly categorized into three main approaches: content-based detection, social-context-based detection, and LLM-based detection.

Content-based approaches focus on analyzing various news content, such as headlines, body text, images, videos, and audio. Language-based methods use deep neural networks to extract linguistic features that distinguish true and fake news through stylistic disparities [21], though such boundaries are increasingly blurred by advanced editorial practices and LLM/AIGC technologies. Some studies enhance linguistic feature via multi-task learning, such as MTEFN [22], which jointly optimizes emotion classification and fake news detection by leveraging emotion-related cues. The second category comprises knowledge-enhanced methods that integrate external knowledge related to news content. For instance, Hu et al. [11] built a topic-entity-sentence heterogeneous graph using knowledge graphs, Yang et al. [23] applied cross-domain transfer to address domain shift, Wang et al. [24] fused generated and original comments to implicitly incorporate expert knowledge. Multi-modal fusion methods combine textual, visual, and auditory data, evolving from early text-image combinations [25–27] to advanced mechanisms such as residual-aware compensation [28], hyperbolic representation framework [29], multi-expert optimization [30], multi-granularity clue alignment [31], and interactive gating for modality interaction [32].

Social-context-based detection techniques include stance-based and propagation-based methods. Stance-based approaches model user reactions, such as attitudes, skepticism, or positional tendencies, to aid detection. For example, Ayooobi et al. [33] proposed ESAS, a metric quantifying user skepticism toward LLM-generated fake news, while HSA-BLSTM [34] uses a hierarchical attention network to extract stance features from posts and sub-events within social contexts. Propagation-based methods instead model dissemination patterns, such as structures, dynamics, or network topologies. Xu et al. [35] proposed NAGASIL which simulates fake news spread

to identify optimal debunkers. Cui et al. [12] proposed RAGCL which encodes rumor propagation tree features via adaptive graph contrastive learning. Zhang et al. [36] introduced BLC to handle uncertainty and noisy relations, while Kim et al. [37] proposed DAWN, which weights edges using engagement timeliness to refine propagation modeling. Sun et al. [38] addressed social context confounders through adversarial multi-debiasing, and Chen et al. [39] developed ReTIP, retrieving relevant participants and their interaction contexts to capture the role of social actors in dissemination.

With the rise of large language models (LLMs), LLM-based fake-news detection has become a new paradigm. They leverage pre-trained knowledge for deep semantic understanding and factual verification, even with limited data. From the perspective of data augmentation and style adaptation, Park et al. [40] proposed adversarial style augmentation to generate diverse stylistic variants of fake news, while Nan et al. [41] generated simulated user comments to supplement scarce early-stage data. For explainability, Wang et al. [42] used LLMs to produce human-readable reasoning chains, and Liu et al. [43] introduced TruEDebate, an LLM-driven multi-agent debate framework enhancing both explainability and detection performance. In knowledge retrieval and generative assistance, Zhou et al. [13] combined LLMs for retrieving up-to-date knowledge with small language models for collaborative detection, Zhang et al. [44] analyzed LLM hallucination patterns to identify LLM-generated fake news, and Hu et al. [45] proposed ARG which employs dual cross-attention to fuse news content with multi-perspective analyses generated by LLMs.

2.2 Deep Clustering

Clustering, as a typical unsupervised learning technique, aims to partition data samples into groups (clusters) based on similarity metrics. However, traditional clustering methods struggle with high-dimensional data, prompting the development of deep clustering approaches.

For instance, McConville et al. [46] proposed a method combining autoencoder embeddings with local manifold learning. This approach preserves autoencoder representational capacity while leveraging the UMAP algorithm [47] to uncover local geometric structures in latent spaces, followed by Gaussian mixture models for clustering. Van Gansbeke et al. [48] proposed SCAN which employs a two-stage framework for unsupervised image classification: It first learns semantic-invariant features via self-supervised learning (e.g., instance discrimination), then optimizes cluster assignments through nearest-neighbor prior-based clustering loss with consistency constraints and entropy regularization. Liu et al. [49] proposed ELCRec which addresses scalability in recommendation systems by integrating user behavior representation learning with end-to-end learnable clustering. This method initializes cluster centers as learnable neurons, designs learnable clustering loss functions to disentangle user intents via cluster separation, and enhances representation-clustering synergy through intent-aware contrastive learning.

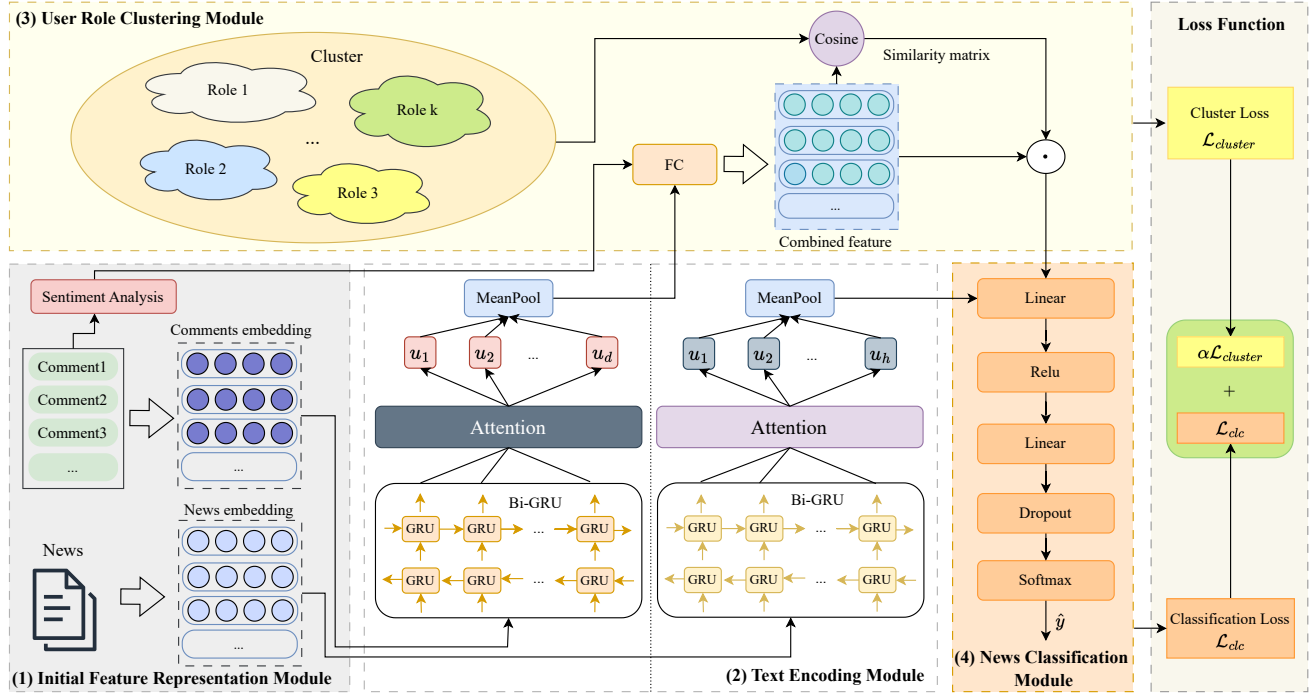


Figure 2: The proposed SARC framework which consists of four components: (1) Initial Feature Representation Module, (2) Text encoding module, (3) User Role Clustering Module, (4) News Classification Module.

3 Problem Statement

Given a news article $A = \{w_i\}_{i=1}^N$ (containing a word sequence of length N) and its associated comment set $C = \{c_j\}_{j=1}^T$ (with T user comments), where each comment c_j is represented as $c_j = \{w_l^j\}_{l=1}^{L_j}$ with w_l^j denoting the l -th word in the j -th comment and L_j denoting the comment length, the objective of fake news detection is to predict the veracity label $\hat{y} = f(A, C; \Theta)$ through a detection model $f(\cdot)$ parameterized by Θ .

4 Methodology

In this section, we will introduce in detail the proposed SARC framework, as shown in Figure 2, which consists of four logically connected components: (1) the Initial Feature Representation Module; (2) the Text Encoding Module; (3) the User Role Clustering Module; and (4) the News Classification Module.

First, SARC employs the Initial Feature Representation Module to generate the feature representations of the news content and user comments, obtaining the sentiment polarity representation of the comments. Subsequently, the Text Encoding Module is used to extract the deep semantic features of the news and the sequential features of the comments respectively. Then, the comment features are enhanced with emotional features. Furthermore, the User Role Clustering Module is adopted to dynamically divide user roles for the sentiment-enhanced comment features in the latent space. Finally, the news semantic vectors and the user role clustering features are fused, and the classification of fake news is achieved through the News Classification Module.

4.1 Initial Feature Representation Module

The initial feature representation module processes news articles and user comments to obtain corresponding embeddings and sentiment polarity representations.

4.1.1 News Embedding. The objective is to convert news text from natural language into vector representations. As news articles typically consist of long texts containing event details such as time, location, and causality, we initialize the embedding layer using word vectors pretrained on large corpora to better capture semantic information. For a news article $A = \{w_i\}_{i=1}^N$ composed of N words, the embedding layer initialized with pretrained word vectors generates word embeddings $T_A = \{t_i\}_{i=1}^N$, where t_i represents the word vector corresponding to w_i .

4.1.2 Comment Embedding. Sharing the same objective of converting text into machine-understandable vector forms, comment embeddings employ separate embedding layers from news embeddings. This design addresses the divergent focuses between news content (emphasizing factual elements like institutional names, locations, and temporal markers) and user comments (dominated by opinions, emotions, and internet slang). For the comment set $C = \{c_j\}_{j=1}^T$ associated with the news, each comment c_j is encoded through a pretrained embedding layer as $T_{c_j} = \{t_l^j\}_{l=1}^{L_{c_j}}$, where t_l^j denotes the word vector corresponding to w_l^j .

4.1.3 Sentiment Analysis. As emotions drive human behavior and fake news demonstrates higher emotional intensity, stronger negativity, and weaker positivity compared to true news [15], user

comments often reflect emotional responses [50]. We employ a pretrained sentiment analysis model to assign ternary sentiment polarity scores (negative/neutral/positive) to comments. The comment set $C = \{c_j\}_{j=1}^T$ is thereby transformed into sentiment polarity representations $E = \{e_j\}_{j=1}^T$, where e_j corresponds to the sentiment polarity of comment c_j .

4.2 Text Encoding Module

Fake news is deliberately crafted to disseminate false information, often employing sensational language to attract attention and generate traffic, resulting in distinct stylistic differences from true news [21]. To capture these disparities, we employ two text encoders to learn semantic representations for news content and user comments respectively.

We utilize Gated Recurrent Units (GRU) [51] to model long-range dependencies and maintain persistent memory. Bidirectional GRU [52] further enhances context awareness by encoding text sequences from both directions.

Given word embeddings $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^L$ for a text sequence (news or comment) of length L , the forward GRU ($\overrightarrow{\text{GRU}}$) and backward GRU ($\overleftarrow{\text{GRU}}$) compute hidden states as:

$$\begin{aligned}\overrightarrow{\mathbf{h}}_i &= \overrightarrow{\text{GRU}}(\mathbf{t}_i, \overrightarrow{\mathbf{h}}_{i-1}), \quad \forall i \in [1, L] \\ \overleftarrow{\mathbf{h}}_i &= \overleftarrow{\text{GRU}}(\mathbf{t}_i, \overleftarrow{\mathbf{h}}_{i+1}), \quad \forall i \in [L, 1]\end{aligned}\quad (1)$$

The bidirectional context-aware representation is obtained through concatenation: $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{h}}_i]$, which encapsulates full-sentence semantics centered at word w_i . To amplify intentionally crafted words (e.g., exaggerations) in fake news, we introduce self-attention [53]. Given GRU hidden states $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L] \in \mathbb{R}^{L \times 2d_h}$ (d_h : GRU hidden dimension), the attention matrix $\mathbf{A} \in \mathbb{R}^{L \times L}$ is computed via scaled dot-product:

$$\mathbf{A}_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_k})}{\sum_{k=1}^L \exp(\mathbf{q}_i^\top \mathbf{k}_k / \sqrt{d_k})}, \quad \begin{cases} \mathbf{q}_i = \mathbf{W}_Q \mathbf{h}_i \\ \mathbf{k}_j = \mathbf{W}_K \mathbf{h}_j \end{cases} \quad (2)$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{2d_h \times d_k}$ are learnable projection matrices. The enhanced representation combines attention weights with residual connections:

$$\mathbf{u}_i = \sum_{j=1}^L \mathbf{A}_{ij} \mathbf{h}_j + \mathbf{h}_i \quad (3)$$

Mean pooling aggregates sequence-level features into fixed-dimensional vectors:

$$\mathbf{v} = \frac{1}{L} \sum_{i=1}^L \mathbf{u}_i, \quad \mathbf{v} \in \mathbb{R}^{2d_h} \quad (4)$$

Given the semantic divergence between news content and user comments, we implement dual encoders with shared architecture but independent parameters:

$$\begin{aligned}\mathbf{v}_A &= \text{encode}_A(\mathbf{T}_A) \\ &= \text{MeanPool}(\text{SelfAttn}(\text{BiGRU}(\mathbf{T}_A))) \\ \mathbf{V}_C &= [\mathbf{v}_{c_1}, \dots, \mathbf{v}_{c_T}] \\ &= [\text{encode}_C(\mathbf{T}_{c_1}), \dots, \text{encode}_C(\mathbf{T}_{c_T})]\end{aligned} \quad (5)$$

4.3 Dynamic Role Clustering Module

During fake news dissemination, user comments often exhibit role differentiation (e.g., spammers, skeptics, blind followers), which differs from true news propagation. To exploit this phenomenon, we employ end-to-end deep clustering algorithms for user role categorization, addressing the limitations of traditional clustering methods in high-dimensional data processing.

Given user comment features $\mathbf{V}_C$ and sentiment features E , we first concatenate the two features:

$$\begin{aligned}\mathbf{V}_C^e &= [\mathbf{v}_{c_1}^e, \mathbf{v}_{c_2}^e, \dots, \mathbf{v}_{c_T}^e] \\ \mathbf{v}_{c_j}^e &= [\mathbf{v}_{c_j}, e_j]\end{aligned} \quad (6)$$

where $\mathbf{v}_{c_j} \in \mathbf{V}_C$, $e_j \in E$.

These features are projected into clustering space via fully connected layer:

$$\mathbf{X} = \text{FC}(\mathbf{V}_C^e) \quad (7)$$

where $\mathbf{X} \in \mathbb{R}^{T \times d_p}$, FC denotes a fully connected layer, T denotes comment count and d_p denotes the clustering space dimension.

We define K trainable cluster centers $\mathbf{M} = [\mu_1, \mu_2, \dots, \mu_K] \in \mathbb{R}^{K \times d_p}$ and apply L2 normalization:

$$\begin{aligned}\tilde{\mathbf{X}} &= \text{L2Norm}(\mathbf{X}, \text{dim} = -1) \\ \tilde{\mathbf{M}} &= \text{L2Norm}(\mathbf{M}, \text{dim} = -1)\end{aligned} \quad (8)$$

We use cosine similarity as the distance between the feature and the cluster center. The cosine similarity matrix $\mathbf{S} = [s_{jk}]_{1 \leq j \leq T}^{1 \leq k \leq K} \in \mathbb{R}^{T \times K}$ is computed as:

$$s_{jk} = \tilde{\mathbf{x}}_j^\top \tilde{\mu}_k \in [-1, 1] \quad (9)$$

Where $\tilde{x}_j$ and $\tilde{\mu}_k$ denote the j -th row of $\tilde{\mathbf{X}}$ and the k -th row of $\tilde{\mathbf{M}}$.

Temperature-scaled soft assignment probabilities are computed as:

$$q_{jk} = \frac{\exp(s_{jk} \times \tau)}{\sum_{k'=1}^K \exp(s_{jk'} \times \tau)} \quad (10)$$

where τ is a learned temperature coefficient. This yields the soft assignment matrix $\mathbf{Q} = [q_{jk}]_{1 \leq j \leq T}^{1 \leq k \leq K} \in \mathbb{R}^{T \times K}$.

Cluster-level semantic representations are aggregated through weighted summation:

$$\mathbf{C}_K = \mathbf{Q}^\top \tilde{\mathbf{X}} \quad (11)$$

where $\mathbf{C}_K \in \mathbb{R}^{K \times d_p}$ contains global semantic features for each role category.

4.4 News Classification Module

This module detects fake news by combining the semantic features of news content with user role features. The two feature sets are merged using a linear transformation and then passed to a classifier.

Given the news semantic feature vector $\mathbf{v}_A \in \mathbb{R}^{2d_A}$ from Eq. (5) and user role cluster feature matrix $\mathbf{C}_K \in \mathbb{R}^{K \times d_p}$ from Eq. (11), fusing news semantic features and user role features through flattening and concatenation:

$$\mathbf{f} = [\mathbf{v}_A; \text{vec}(\mathbf{C}_K)] \in \mathbb{R}^{2d_A + K \cdot d_p} \quad (12)$$

The classification decision is realized through deep nonlinear mapping:

$$\begin{aligned}\mathbf{h}_1 &= \text{Dropout}(\sigma(\mathbf{W}_1 \mathbf{f} + \mathbf{b}_1)) \\ \hat{y} &= \text{Softmax}(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2)\end{aligned} \quad (13)$$

where σ denotes the ReLU activation function.

4.5 Loss Function

The model achieves collaborative learning of fake news detection and user role cluster through multi-task joint optimization. The total loss combines supervised classification loss and unsupervised clustering loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{cluster}} \quad (14)$$

where α balances the importance of unsupervised clustering loss.

4.5.1 Classification Loss. We employ L2-regularized cross-entropy loss for each sample:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B} \sum_{b=1}^B \sum_{c=1}^M y_{bc} \log(p_{bc}) \quad (15)$$

where y_{bc} denotes the true class label and p_{bc} is the predicted probability for class c .

4.5.2 Unsupervised Clustering Loss. The clustering loss consists of intra-class loss $\mathcal{L}_{\text{intra}}$ and inter-class loss $\mathcal{L}_{\text{inter}}$:

$$\mathcal{L}_{\text{cluster}} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}} \quad (16)$$

To prevent early-stage errors from being reinforced [49], the intra-class loss is designed to pull features toward all cluster centers:

$$\mathcal{L}_{\text{intra}} = -\frac{1}{B \cdot T} \sum_{b=1}^B \sum_{j=1}^T \sum_{k=1}^K q_{bjk} \cdot \underbrace{\tilde{\mathbf{x}}_{bj}^T \tilde{\mu}_k}_{\text{cosine similarity}} \quad (17)$$

The inter-class loss separates cluster centers:

$$\mathcal{L}_{\text{inter}} = \frac{1}{K(K-1)} \sum_{k=1}^K \sum_{\substack{k'=1 \\ k' \neq k}}^K \tilde{\mu}_k^T \tilde{\mu}_{k'} \quad (18)$$

Here $q_{bjk} \in [0, 1]$ represents the probability of comment j in batch b being assigned to cluster k , with $\tilde{\mathbf{x}}_{bj} = \frac{\mathbf{x}_{bj}}{\|\mathbf{x}_{bj}\|_2}$ and $\tilde{\mu}_k = \frac{\mu_k}{\|\mu_k\|_2}$ being L2-normalized features and cluster centers respectively.

5 Experimental Evaluation

5.1 Datasets

To evaluate the performance of SARC in fake news detection, we conduct experiments on the Chinese dataset Weibo-comp [54] and the English dataset RumourEval-19 [55].

5.1.1 Weibo-comp. It is a multimodal dataset containing Weibo posts with original texts, comments, images, and domain categories. It covers news content from eight domains (e.g., technology, politics, military, etc.) on Weibo platform, including COVID-19-related information. The dataset contains three categories: real, fake, and undecidable news. Here, we use real and fake news with corresponding comments for experiments.

5.1.2 RumourEval-19. It originates from the SemEval-2019 rumor detection task, containing news posts from Twitter and Reddit. It categorizes news into three classes: real, fake, and unverified. Detailed dataset statistics are shown in Table 1.

Metric	RumourEval-19	Weibo-comp
Total News	445	6,444
Real News	138	2,863
Fake News	185	3,581
Unverified News	122	–
Comments Count	7,990	102,372
Avg. Comments per News	18	16

Table 1: The statistics of RumourEval-19 and weibo-comp

5.2 Baseline Methods

To comprehensively evaluate the performance of SARC, we compare it with representative baseline models, which include: basic text learning models LSTM, Text-CNN, BERT, and HAN; social context-based models, namely HSA-BLSTM and dEFEND; sentiment-based models Dual-Emo and MTEFN; and models integrating the capabilities of Large Language Models (LLMs), including BERT w/G, dEFEND w/G, ARG, and ARG-D. Here w/G means the models are supplied with comment features generated by GenFEND [41].

- **LSTM** [56]: A classic sequential model that processes word sequence features through average pooling of hidden states, followed by a fully connected layer for prediction.
- **Text-CNN** [57]: A CNN-based text classifier which uses convolution kernels of varying sizes to capture multi-granular semantic patterns for fake news detection.
- **BERT** [58]: Transformer-based pretrained language model fine-tuned with the [CLS] token’s hidden state as news representation.
- **BERT w/G**: It introduces the comment features generated by GenFEND [41] to the BERT model, forming a joint input of news content features and generated comment features to enhance detection capability.
- **HAN** [59]: Hierarchical attention network with bi-GRU layers modeling word-level and sentence-level features through paragraph segmentation.
- **HSA-BLSTM** [34]: Hierarchical attention network extracting features from words, posts, and sub-events across news and social contexts.
- **dEFEND** [14]: Co-attentive framework combining news sentences and user comments through sentence-comment co-attention mechanism.
- **dEFEND w/G** [14]: It supplements the dEFEND model with the comment features generated by GenFEND [41], forming a joint input of news content features, actual comment features, and generated comment features.
- **Dual-Emo** [17]: HSA-BLSTM enhanced with publisher-user bidirectional emotional features for emotion-aware detection.
- **MTEFN** [22]: Multi-task framework jointly optimizing emotion classification and fake news detection.
- **ARG** [45]: ARG realizes the feature interaction between news text and multi-perspective analyses generated by LLMs through a dual cross-attention mechanism, dynamically screens

and fuses useful information, and finally generates fake news predictions through aggregated features.

- **ARG-D** [45]: ARG-D is a distilled version of ARG, which reuses the news encoder and classifier parameters of ARG to realize fake news detection without querying LLMs.

Parameter	Weibo-comp	RumourEval-19
Embedding dimension	300	300
News encoding hidden dim	256	256
Comment encoding hidden dim	128	128
Comment aggregation FC dim	256	256
Comment emotion feature dim	1	1
Number of clusters	3	3
Batch size	8	8
Learning rate	1e-3	5e-4
Optimizer		Adam
Dropout rate		0.5
Training epochs		20

Table 2: Hyperparameter setting

For the RumourEval-19 dataset, we follow the official competition metrics: Macro-F1 and RMSE. For Weibo-comp, we adopt Accuracy, Precision, Recall, and F1-score as evaluation metrics.

5.3 Implementation Details

On RumourEval-19 and Weibo-comp, the datasets are split into 7:1:2 and 6:2:2 ratios for training, validation, and testing, respectively. As shown in Table 2, the embedding dimension is fixed at 300, with hidden dimensions of 256 for the news encoder and 128 for the comment encoder. The comment aggregation FC dimension is 256, and each comment’s emotion feature is represented by a single scalar. The role clustering module uses 3 clusters for both datasets.

For the Weibo-comp dataset, we extract sentiment features using a fine-tuned BERT model and perform Chinese word segmentation with Jieba; the embedding layer is initialized using Chinese Word Vectors [60, 61]. For RumourEval-19, sentiment features are obtained via VaderSentiment [62], tokenization is carried out using NLTK, and the embedding layer is initialized with GloVe embeddings [63, 64].

Training uses a batch size of 8 and the Adam optimizer, with learning rates of 1×10^{-3} for Weibo-comp and 5×10^{-4} for RumourEval-19. A dropout rate of 0.5 is applied, and training proceeds for 20 epochs. These hyperparameters are selected based on preliminary experiments to balance performance and efficiency.

5.4 Overall Results

The experimental results are shown in Table 3, where the best results are shown in bold while the second best ones are underlined. It is observed that SARC significantly outperforms other comparative methods in all classification metrics. On the RumourEval-19 dataset, the Macro-F1 of SARC is 0.357, which is significantly higher than that of the existing best-performing Dual-BLSTM (Macro-F1 = 0.334). The RMSE value of SARC is 0.761, which is significantly lower than that of all comparative models. This indicates that SARC can also perform well when the dataset is small.

Method	RumourEval-19		Weibo-comp			
	Macro-F1	RMSE	Accuracy	Precision	Recall	F1-score
LSTM	0.254	1.186	0.816	0.826	0.849	0.837
Text-CNN	0.267	1.042	0.829	0.809	0.906	0.855
HAN	0.277	0.901	0.846	0.845	0.885	0.865
BERT	0.294	0.988	0.884	0.885	0.910	0.897
BERT w/G	0.327	0.855	0.872	0.875	0.881	0.878
HSA-BLSTM	0.310	0.975	0.890	0.893	0.912	0.902
Dual-Emo	0.334	0.824	0.903	<u>0.921</u>	0.902	0.911
dEFEND	0.327	1.136	0.901	0.914	0.908	0.911
dEFEND w/G	0.315	<u>0.816</u>	<u>0.920</u>	0.918	<u>0.922</u>	<u>0.920</u>
MTEFN	0.321	0.879	0.835	0.880	0.815	0.846
ARG	0.294	0.850	0.911	0.911	0.903	0.907
ARG-D	0.287	0.868	0.872	0.881	0.868	0.874
SARC	0.357	0.761	0.969	0.967	0.967	0.967

Table 3: Performance comparison between SARC and baselines

For the Weibo-comp dataset, SARC also achieves the best performance. Even compared with the second best results, SARC still exhibits remarkable improvements of 4.9%, 4.6%, 4.5%, and 4.7% in terms of Accuracy, Precision, Recall, and F1-score, respectively. The accuracy improvement of SARC over the baseline models ranges from 4.9% to 15.3%, proving its outstanding performance in detecting real and fake news. To sum up, the combination of the semantic features of news content and user role features by SARC breaks the limitations brought by simple sentiment analysis and text modeling, and achieves better experimental results.

5.5 Ablation Studies

We conduct four ablation studies to evaluate each module’s contribution. First, SARC-n removes the news encoder, relying solely on the dynamic role clustering module. Second, SARC-c eliminates the dynamic role clustering module while retaining other components. Third, SARC-e discards comment sentiment features. Finally, SARC-l only uses the classification loss and discards the clustering loss. Results in Table 3 and visualizations in Figure 3 and Figure 4 demonstrate each module’s impact.

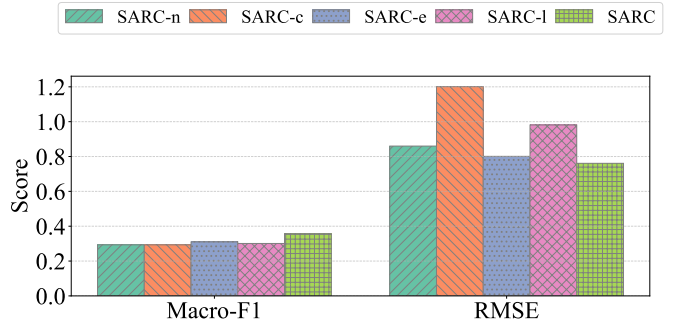


Figure 3: Performance comparison of SARC and its variant models on the RumourEval-19 dataset.

Ablation results demonstrate the synergistic effects of SARC modules. As shown in Table 4, removing the news encoder (SARC-n) causes Macro-F1 on RumourEval-19 to plummet from 0.357 to 0.293,

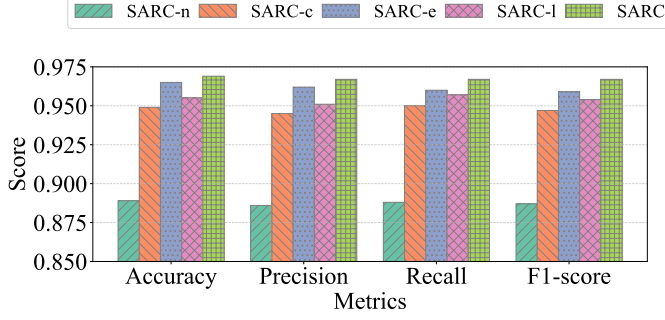


Figure 4: Performance comparison of SARC and its variant models on the Weibo-comp dataset.

Method	RumourEval-19		Weibo-comp			
	Macro-F1	RMSE	Accuracy	Precision	Recall	F1-score
SARC-n	0.293	0.860	0.889	0.886	0.888	0.887
SARC-c	0.293	1.201	0.949	0.945	0.950	0.947
SARC-e	0.310	0.801	0.965	0.962	0.960	0.959
SARC-l	0.301	0.981	0.955	0.951	0.957	0.954
SARC	0.357	0.761	0.969	0.967	0.967	0.967

Table 4: Ablation results of SARC

confirming this module’s critical role in extracting core semantic features. Disabling the dynamic role clustering module (SARC-c) reduces Weibo-comp’s F1-score by 1.3% to 0.954, proving its effectiveness in enhancing detection through comment semantic clustering. Eliminating sentiment features (SARC-e) increases RMSE by 4% on RumourEval-19, highlighting emotional signals’ contribution. SARC-l, which only uses classification loss, also shows a decline in performance, indicating the synergy of joint optimization of classification and clustering. The progressive performance degradation (SARC-n < SARC-c < SARC-l < SARC-e < SARC) reveals module complementarity, with the full SARC achieving optimal metrics across datasets, thereby validating the combined efficacy of semantic modeling, clustering, and sentiment analysis.

5.6 Hyperparameter Sensitivity Analysis

We conducted two hyperparameter sensitivity analysis experiments, adjusting the number of clusters $k \in \{2, 3, 4, 5, 10, 20, 50, 100\}$ and the value of the clustering loss $\alpha \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$ in the loss function respectively. The experimental results are shown in Figure 5.

As can be seen from Figure 5, the number of clusters k has a certain impact on the model performance, especially when k lies within the range of $[1, 10]$, and the model performs best when $k=3$. On RumourEval-19, the performance of the model decreases as k increases, but the degree of decrease is insignificant when k becomes too large. On the Weibo dataset, the model’s performance tends to be stable when $k \geq 10$.

We conjecture this phenomenon is mainly due to the setting of primary and secondary weights of the loss function during training. Specifically, the clustering module is introduced as an auxiliary task in this model, and its loss is not the core target of training

optimization. Though the value of k changes, the classification loss still dominates, so the overall target optimization direction of the model will not shift drastically.

Moreover, because the clustering module adopts a soft assignment mechanism, the model can automatically adjust the correlation strength between samples and clusters, weakening the hard constraint of k . Consequently, the model possibly only activates a small number of effective clusters. As shown in Figure 6, when the number of clusters is 4 or more, the sample size of some clusters is far lower than that of other clusters. This indicates that the model does not treat all clusters equally, but tends to focus on cluster structures with more concentrated information.

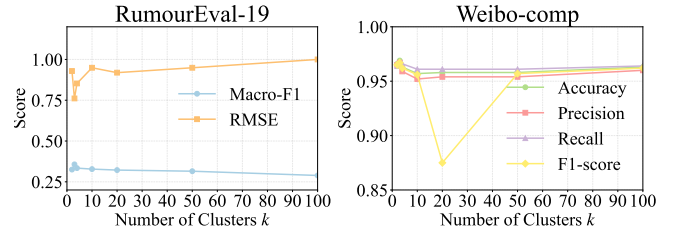


Figure 5: The impact of parameter k on model performance.

Table 7 shows the model performance under different settings of the clustering loss weight α . The experimental results indicate that the model is somewhat sensitive to this hyperparameter: an excessively small or large α will both lead to performance degradation, while a moderate range can significantly improve the model’s detection performance. On RumourEval-19, the model performance generally declines as α increases and eventually plateaus. This behavior likely stems from the dataset’s small size: when the clustering-loss weight becomes too large, the model quickly bottoms out. Weibo, being much larger, exhibits a more linear sensitivity to α , where the performance metrics keep sliding without any sign of leveling off as α continues to grow.

Specifically, when $\alpha = 0.05$, the model achieves optimal performance. On RumourEval-19, the Macro-F1 increases to 0.357 and the RMSE decreases to 0.761; on Weibo-comp, the Accuracy and F1-score reach 0.967 respectively, which is significantly better than other parameter settings. This suggests that under this weight, the clustering loss and classification loss achieve a well-balanced collaborative optimization.

When α is set too small (e.g., 0.01), the clustering module fails to provide sufficient guidance for the overall training, making it difficult to effectively exert its auxiliary function, resulting in only a slight improvement in classification performance. Conversely, when α takes a larger value (e.g., $\alpha \geq 0.1$), the clustering loss accounts for an excessively high proportion in training, which instead inhibits the classification learning process of the main task, and the model tends to suffer from the problem of optimization direction deviation. For example, when $\alpha = 2$, the Macro-F1 on RumourEval-19 drops to 0.289, and the Accuracy on Weibo-comp drops to 0.932, showing obvious performance degradation.

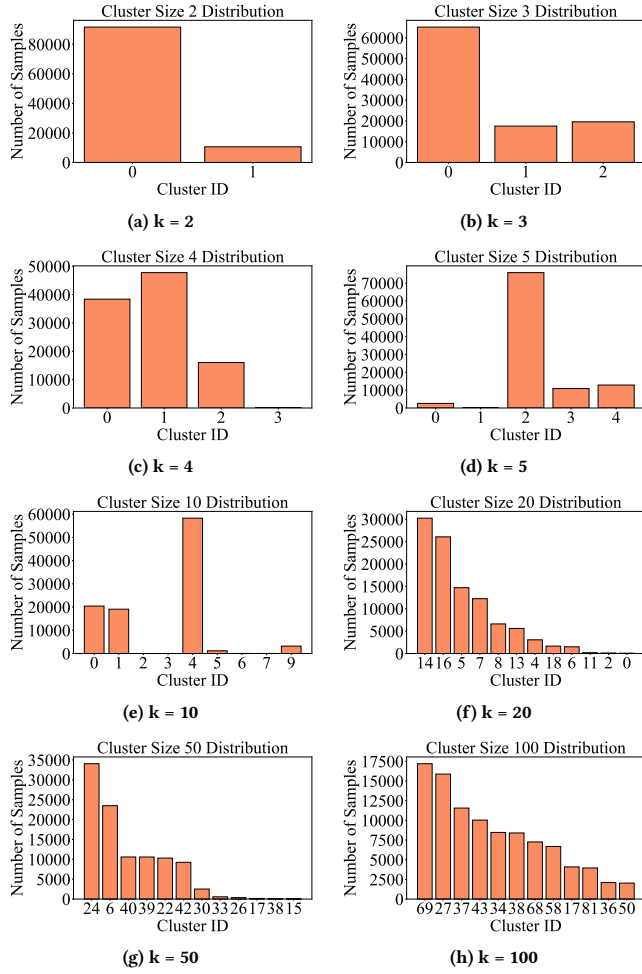


Figure 6: Visualization of clustering results with different numbers of clusters

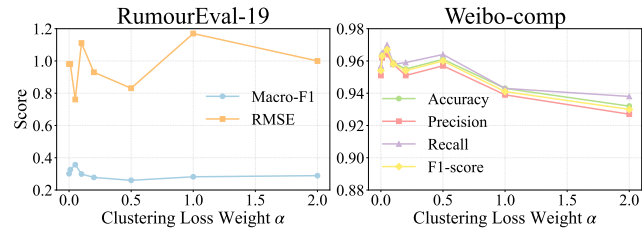


Figure 7: The impact of parameter α on model performance

5.7 Visual Analysis

Figure 8 visualizes feature distributions of the Weibo dataset using t-SNE-based 3D projections with cluster-specific color coding, where Figure 8(a) displays fake news clusters ($n=48,570$), while Figure 8(b) shows true news distributions ($n=53,656$). Three clusters emerge: green (central), yellow (lower-left), and blue (upper-right), demonstrating clear cluster boundaries and intra-cluster compactness.

Notably, fake news exhibits significantly more yellow points than true news, revealing comment content divergence and validating the dynamic clustering module's effectiveness.

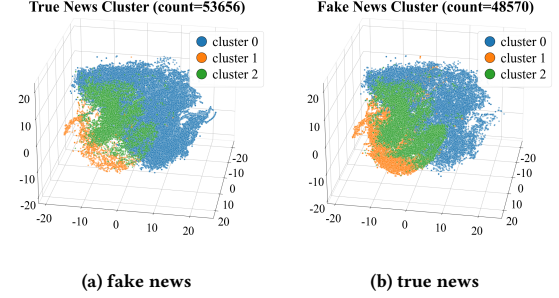


Figure 8: Visualization results of features on Weibo-comp dataset

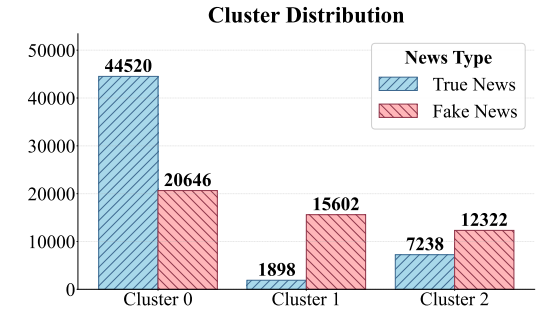


Figure 9: Cluster distribution on Weibo-comp dataset

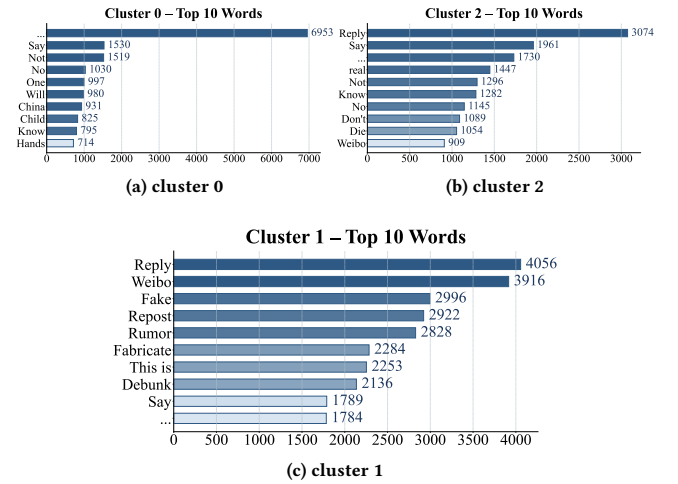


Figure 10: Word frequency statistics of clustering clusters

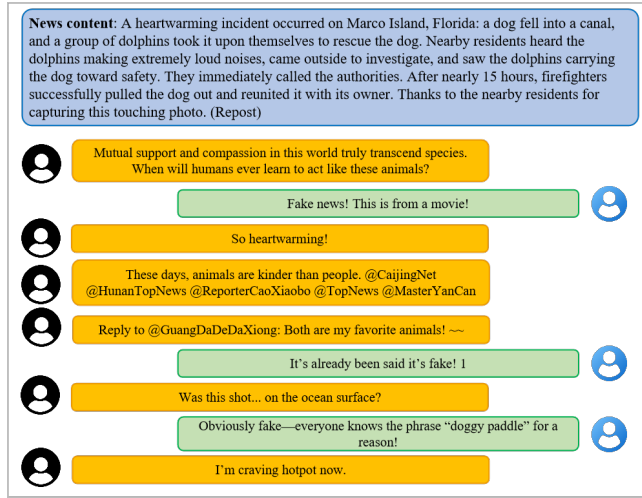


Figure 11: News example of cluster 1

Figure 9 quantifies cluster distribution differences. True news concentrates in Cluster 0 and Cluster 2 (Cluster 0 dominates with $n=42,520$, 86.05%) reflecting semantic consistency. Fake news shows balanced inter-cluster distribution, with Cluster 1 comments being 8 times more prevalent. Word frequency analysis (Figure 10) reveals Cluster 1’s top terms, from which we see “fake”, “rumor”, “Fabricate” and “Debunk” ranked 3rd, 5th, 6th, and 8th respectively. Figure 11 shows a real-world fake news post and its comments. Cluster 1 (in green) mainly contains debunking and skeptical remarks, confirming Cluster 1 primarily hosts debunking comments in fake news. This is consistent with the result in Figure 9 that the number of Cluster 1 differs by 8 times, indicating a positive correlation between debunking behaviors and fake news.

6 Conclusion

In this paper, we tackled the fake news detection problem from a user role perspective using deep clustering. Learnable cluster centers dynamically group user comments, improving both behavior understanding and model explainability. Experiments on multiple datasets exhibited the model’s outstanding performance, and ablation studies confirmed each module’s positive impact.

However, there remains room for improvement. First, the current clustering algorithm uses a fixed number of clusters, which may fail to adequately capture potentially important user roles in the data. Second, the BiGRU architecture combined with an attention mechanism may encounter performance bottlenecks when dealing with large-scale comment data. Future work could introduce deep clustering methods with adaptive cluster numbers and, when facing an overwhelming volume of comments, prioritize high-impact ones based on engagement metrics such as likes and shares to improve efficiency. Additionally, more effective strategies for fusing sentiment and textual features could be explored, along with the incorporation of multi-modal data. Furthermore, user-related metadata—such as account age and posting frequency—could be integrated into the role differentiation module to further enhance model performance.

7 Ethical Considerations

The proposed SARC framework, while aiming to improve fake news detection through sentiment-augmented role clustering, presents several ethical challenges. Biases in training data may propagate into the learned user role clusters and detection results, potentially reinforcing stereotypes or unfairly targeting specific groups. Since user comments often reflect cultural norms and linguistic nuances, a model trained in one context may misclassify in another, raising cross-cultural fairness concerns. Overly aggressive detection thresholds risk suppressing legitimate speech, especially from marginalized voices or those critical of authority, while false negatives could enable harmful misinformation to spread. The unsupervised nature of the clustering process may limit explainability, making it more difficult to identify and correct embedded biases. Moreover, such technology could be misused to silence dissent under the pretext of combating “fake news”.

References

- [1] Daniele Notarmuzi, Claudio Castellano, Alessandro Flammini, Dario Mazzilli, and Filippo Radicchi. 2022. Universality, criticality and complexity of information propagation in social media. *Nature communications* 13, 1 (2022), 1308.
- [2] Jiaxing Shang, Xueqi Jia, Xiaoquan Li, Fei Hao, Ruiyuan Li, and Geyong Min. 2025. DVCAE: Semi-Supervised Dual Variational Cascade Autoencoders for Information Popularity Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2025).
- [3] Haoyu Xiong, Jiaxing Shang, Fei Hao, Dajiang Liu, and Geyong Min. 2025. SDVD: Self-supervised dual-view modeling of user and cascade dynamics for information diffusion prediction. *Knowledge-Based Systems* (2025), 114005.
- [4] Akhil Bhimaraju, Eliot W Robson, Lav R Varshney, and Abhishek K Umrawal. 2024. Fractional Budget Allocation for Influence Maximization under General Marketing Strategies. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 3627–3631.
- [5] Pritish Chakraborty, Sayan Ranu, Krishna Sri Ipsit Mantri, and Abir De. 2023. Learning and maximizing influence in social networks under capacity constraints. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 733–741.
- [6] Chidiogo Uzoamaka Akpuokwe, Chidinma Favour Chikwe, and Nkechi Emanuella Eneh. 2024. Innovating business practices: The impact of social media on fostering gender equality and empowering women entrepreneurs. *Magna Scientia Advanced Research and Reviews* 10, 2 (2024), 032–043.
- [7] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.
- [8] Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers* 26, 2 (2024), 443–458.
- [9] Jiafeng Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep’s clothing: Robust fake news detection against LLM-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 3367–3378.
- [10] Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 435–445.
- [11] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*. Association for Computational Linguistics, Online, 754–763.
- [12] Chaoyun 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*, Vol. 38. AAAI Press, 73–81.
- [13] Ziyi Zhou, Xiaoming Zhang, Shenghan Tan, Litian Zhang, and Chaozhao Li. 2025. Collaborative evolution: Multi-round learning between large and small language models for emergent fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, 1210–1218.
- [14] 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*. Association for

- Computing Machinery, New York, NY, USA, 395–405.
- [15] Lina Zhou, Jie Tao, and Dongsong Zhang. 2023. Does fake news in different languages tell the same story? An analysis of multi-level thematic and emotional characteristics of news about COVID-19. *Information Systems Frontiers* 25, 2 (2023), 493–512.
 - [16] Kailun Zhu, Songtao Peng, Jiaqi Nie, Zhongyuan Ruan, Shanqing Yu, and Qi Xuan. 2024. Exploring agent interaction patterns in the comment sections of fake and real news. *Journal of the Royal Society Interface* 21, 220 (2024), 20240483.
 - [17] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*. 3465–3476.
 - [18] Siqi Jiang, Zeqi Guo, and Jihong Ouyang. 2024. What makes sentiment signals work? Sentiment and stance multi-task learning for fake news detection. *Knowledge-Based Systems* 303 (2024), 112395.
 - [19] Hao Zhang, Zonglin Li, Sanya Liu, Tao Huang, Zhouwei Ni, Jian Zhang, and Zhihan Lv. 2023. Do sentence-level sentiment interactions matter? Sentiment mixed heterogeneous network for fake news detection. *IEEE Transactions on Computational Social Systems* 11, 4 (2023), 5090–5100.
 - [20] Ruichao Yang, Jing Ma, Wei Gao, and Hongzhan Lin. 2025. Llm-enhanced multiple instance learning for joint rumor and stance detection with social context information. *ACM Transactions on Intelligent Systems and Technology* 16, 3 (2025), 1–27.
 - [21] 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* (2017).
 - [22] Arjun Choudhry, Inder Khatri, Minni Jain, and Dinesh Kumar Vishwakarma. 2022. An emotion-aware multitask approach to fake news and rumor detection using transfer learning. *IEEE Transactions on Computational Social Systems* 11, 1 (2022), 588–599.
 - [23] Xuankai Yang, Yan Wang, Xiuzhen Zhang, Shoujin Wang, Huaxiong Wang, and Kwok Yan Lam. 2025. A Macro-and Micro-Hierarchical Transfer Learning Framework for Cross-Domain Fake News Detection. In *Proceedings of the ACM on Web Conference 2025*. 5297–5307.
 - [24] Bing Wang, Bingrui Zhao, Ximing Li, Changchun Li, Wanfu Gao, and Shengsheng Wang. 2025. Collaboration and Controversy Among Experts: Rumor Early Detection by Tuning a Comment Generator. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 468–478.
 - [25] Priyanka Meel and Dinesh Kumar Vishwakarma. 2021. HAN, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences* 567 (2021), 23–41.
 - [26] Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhao Li. 2024. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. AAAI Press, 16777–16785.
 - [27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
 - [28] Xinquan Yu, Ziqi Sheng, Wei Lu, Xiangyang Luo, and Jiantao Zhou. 2025. Racmc: Residual-aware compensation network with multi-granularity constraints for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, 986–994.
 - [29] Shanshan Feng, Guoxin Yu, Dawei Liu, Han Hu, Yong Luo, Hui Lin, and Yew-Soon Ong. 2025. MHR: A Multi-Modal Hyperbolic Representation Framework for Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering* (2025).
 - [30] Lingzhi Shen, Yunfei Long, Xiaohao Cai, Imran Razzak, Guanming Chen, Kang Liu, and Shoaib Jameel. 2025. Gamed: Knowledge adaptive multi-experts decoupling for multimodal fake news detection. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 586–595.
 - [31] Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2025. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI Press, 228–236.
 - [32] Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*. Association for Computing Machinery, New York, NY, USA, 5139–5150.
 - [33] Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2024. Seeing Through AI's Lens: Enhancing Human Skepticism Towards LLM-Generated Fake News. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. Association for Computing Machinery, New York, NY, USA, 1–11.
 - [34] Han Guo, Juan Cao, Yazhi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*. Association for Computing Machinery, New York, NY, USA, 943–951.
 - [35] Xiaofei Xu, Ke Deng, Michael Dann, and Xiuzhen Zhang. 2024. Harnessing network effect for fake news mitigation: Selecting debunkers via self-imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, 22447–22456.
 - [36] Guixian Zhang, Shichao Zhang, and Guan Yuan. 2024. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data* 18, 4 (2024), 1–21.
 - [37] Junghoon Kim, Junmo Lee, Yeonjun In, Kanghoon Yoon, and Chanyoung Park. 2025. Revisiting Fake News Detection: Towards Temporality-aware Evaluation by Leveraging Engagement Earliness. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 559–567.
 - [38] Zixun Sun, Mingye Xu, Guanming Liang, and Qi Liu. 2025. Unifying Adversarial Multi-Deconfounded Learning Paradigm for Fake News Detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. Association for Computing Machinery, New York, NY, USA, 2572–2583.
 - [39] Xueqin Chen, Xiaoyu Huang, Qiang Gao, Li Huang, Jiajing Yu, and Guisong Liu. 2025. Birds of a Feather: Enhancing Multimodal Fake News Detection via Multi-Element Retrieval. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE Computer Society, IEEE Computer Society, Los Alamitos, CA, USA, 3615–3628.
 - [40] Sungwon Park, Sungwon Han, Xing Xie, Jae-Gil Lee, and Meeyoung Cha. 2025. Adversarial Style Augmentation via Large Language Model for Robust Fake News Detection. In *Proceedings of the ACM on Web Conference 2025*. 4024–4033.
 - [41] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1732–1742.
 - [42] Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*. 2452–2463.
 - [43] Yuhao Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 504–514.
 - [44] Chaowei Zhang, Zongling Feng, Zewei Zhang, Jipeng Qiang, Guandong Xu, and Yun Li. 2025. Is LLMs Hallucination Usable? LLM-based Negative Reasoning for Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1031–1039.
 - [45] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. AAAI Press, 22105–22113.
 - [46] Ryan McConville, Raul Santos-Rodriguez, Robert J Piechocki, and Ian Craddock. 2021. N2d(not too) deep clustering via the local manifold of an autoencoded embedding. In *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 5145–5152.
 - [47] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
 - [48] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *European conference on computer vision*. Springer, Springer-Verlag, Berlin, Heidelberg, 268–285.
 - [49] Yue Liu, Shihao Zhu, Jun Xia, Yingwei Ma, Jian Ma, Xinwang Liu, Shengju Yu, Kejun Zhang, and Wenliang Zhong. 2024. End-to-end learnable clustering for intent learning in recommendation. *Advances in Neural Information Processing Systems* 37 (2024), 5913–5949.
 - [50] Christy Galletta Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2023. Emotions: The unexplored fuel of fake news on social media. In *Fake News on the Internet*. Routledge, 147–174.
 - [51] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1724–1734.
 - [52] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*. MIT Press, 55 Hayward St., Cambridge, MA, United States.
 - [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
 - [54] Beijing Municipal Bureau of Economy and Information Technology, CCF Big Data Technical Committee, and Institute of Computing Technology, Chinese Academy of Sciences. 2020. Internet False Information Competition Dataset During the COVID-19 Pandemic [in Chinese]. <https://www.datafountain.cn/competitions/422>
 - [55] G Gorrell, K Bontcheva, L Derczynski, E Kochkina, M Liakata, and A Rumourevall Zubiaga. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity

- and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 845–854.
- [56] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [57] Yahui Chen. 2015. *Convolutional neural network for sentence classification*. Master's thesis. University of Waterloo.
- [58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.
- [59] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.
- [60] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Melbourne, Australia). Association for Computational Linguistics, 138–143. <http://aclweb.org/anthology/P18-2023>
- [61] Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting Correlations between Intrinsic and Extrinsic Evaluations of Word Embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 209–221.
- [62] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.
- [63] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. <http://www.aclweb.org/anthology/D14-1162>
- [64] Riley Carlson, John Bauer, and Christopher D Manning. 2025. A New Pair of GloVes. *arXiv preprint arXiv:2507.18103* (2025).