RESEARCH PAPER
Contrib. Sci. & Tech Eng, 2025, 2(4)
Homepage: https://cste.journals.umz.ac.ir/
DOI: 10.22080/cste.2025.29798.1073

Synergistic Content Understanding: Misinformation Detection through Contrastive Regularization and Embedding-Space Mixup

Mojtaba Padashi [1], Meysam Roostaee [1]*, Hassan Zeynali [1], Alireza Jafari [1]

[1] *Department of Computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran.*

Article Info
Received 07 August 2025
Accepted 01 September 2025
Available online 28 September 2025

Keywords:
Misinformation Detection;
Representation Learning;
Supervised Contrastive Learning;
Embedding-Space Mixup.

Abstract:
Automated fake-news detection is a critical challenge for preserving the integrity of the online information ecosystem. Current state-of-the-art systems increasingly depend on external context, such as social propagation graphs, which fundamentally limits their applicability in real-time or "cold-start" scenarios where such signals are unavailable. We challenge the prevailing assumption that this external context is indispensable for top-tier performance. Instead, we argue that the primary bottleneck is the brittle and poorly structured content representations learned via standard model fine-tuning. To address this, we propose a synergistic training framework that sculpts a more robust and discriminative embedding space. Our method harmonizes two complementary and powerful techniques: (1) supervised contrastive regularization, which explicitly structures the feature space by enforcing tight intra-class clustering and clear inter-class separation, and (2) embedding-space mixup, a regularization strategy that creates smoother, more generalizable decision boundaries. On two widely used public benchmarks, Twitter15 and Twitter16, our purely content-only framework establishes a new state-of-the-art, achieving Weighted F1-scores of 94.2% and 94.7%, respectively, and significantly outperforms not only other text-based models but also leading context-aware methods. Our results demonstrate that, with a sufficiently rigorous training regimen, the intrinsic signals within text alone can drive superior veracity assessment.

***Corresponding Author:** m.roostaee@umz.ac.ir

Supplementary information: Supplementary information for this article is available at https://cste.journals.umz.ac.ir/

Please cite this paper as: Padashi, M., Roostaee, M., Zeynali, H., & Jafari, A. (2025). Synergistic Content Understanding: Misinformation Detection through Contrastive Regularization and Embedding-Space Mixup. Contributions of Science and Technology for Engineering, 2(4), 59-68. doi:10.22080/cste.2025.29798.1073

1. Introduction

The proliferation of "fake news", ranging from subtle misinformation to overt propaganda, steadily erodes public trust and endangers societal well-being [1, 2]. Automated detection is imperative, yet current approaches remain constrained. Context-based models, which leverage social diffusion patterns or user metadata [3], excel when rich historical data is available; however, they fail in "cold-start" settings where such signals are absent. Similarly, knowledge-based systems that verify claims against external repositories [4] often lag behind in breaking news or hyper-local events that have not yet been captured in their databases.

This leaves content-only methods as the most immediately deployable paradigm, but they are often dismissed as inherently weaker. We argue that this perceived weakness arises not from a fundamental lack of veracity signals within the text, but rather from brittle representations learned through standard fine-tuning, where minor paraphrases or stylistic adjustments can easily mislead a model [5]. We hypothesize that, with the proper training regimen, the intrinsic signals within text alone are sufficient for state-of-the-art veracity assessment.

To test this hypothesis, we introduce a synergistic training framework that sculpts a resilient embedding space from text by uniting two complementary techniques: (1) supervised contrastive regularization to systematically structure the feature space by clustering same-class examples, and (2) embedding-space mixup to regularize the model and learn smoother, more generalizable decision boundaries.

On two standard public benchmarks (Twitter15 and Twitter16), our framework establishes a new state-of-the-art for content-only models. Critically, it consistently outperforms even leading context-aware systems that rely on external social graph information, demonstrating the profound impact of a superior content representation.

Our contributions are threefold:

ISSN ***3060-6578***

1. A Novel Training Framework: We propose and evaluate the first cohesive integration of supervised contrastive loss and embedding-space mixup specifically for the task of misinformation detection.
2. Revisiting the Role of Context: We provide strong empirical evidence that a content-only model, when trained with a robust representation learning objective, can surpass complex, context-dependent models.
3. Demonstrating Synergy: Through a detailed ablation study, we show that the combination of contrastive learning and mixup yields significant performance gains beyond what either component can achieve alone, proving their synergistic effect.

2. Related Work

The automated detection of misinformation is a vast, complex, and rapidly evolving field, as evidenced by several recent comprehensive reviews [6–8]. Research efforts can be broadly categorized into three dominant research paradigms: context-based, knowledge-based, and content-based approaches. This section situates our proposed framework within this landscape, arguing that by addressing fundamental limitations in content representation, our method establishes a new and more robust foundation for the field. A summary of the key approaches and their limitations is presented in Table 1.

2.1. Context-Aware Methods

A significant body of work is predicated on the premise that a claim's veracity is best determined not by its content alone, but by the external signals related to its origin and propagation.

Social Context Models: This paradigm treats misinformation as a socio-technical phenomenon, analyzing its spread through a network graph. Foundational approaches model the propagation structure of news cascades, using Graph Neural Networks (GNNs) to learn the distinct tree-like patterns of how fake and real news diffuse [9, 10]. More recent work has advanced this frontier by utilizing structural contrastive learning on propagation trees, thereby eliminating the need for labeled data [3]. The SiMiD model [11] operationalized this by using contrastive learning to separate posts based on their community of origin, while models like DANES [12] build ensemble architectures with dedicated social context branches. The primary limitation of these models is their heavy reliance on rich, explicit graph data, which is often unavailable or absent entirely at a post's inception, creating a critical "cold start" problem.

To illustrate, consider a breaking news event, such as an unconfirmed report of an explosion. A malicious actor posts a tweet with a false claim: "Explosion at City Hall confirmed as a terrorist attack." At the moment of its creation (t=0), this post has neither a propagation tree nor user engagement. Consequently, propagation-based models (e.g., Bi-GCN) cannot render a verdict because no diffusion pattern has yet emerged. Similarly, community-based models like SiMiD are also ineffective, as the post has not yet been adopted or amplified by any identifiable community. It is precisely in these first critical minutes and hours that the falsehood spreads most rapidly, often reaching thousands of users before any contextual signals have accumulated. This limitation has significant implications for the design of real-world information systems [13], as it means context-based models cannot provide immediate verdicts.

Table 1. Misinformation detection paradigms and SCU's advantages over each other

Category	Core Idea	Representative Works	SCU's Advantage
Social Context	Models propagation graphs and user interactions	SiMiD [11], SCL [3], DANES [12]	No reliance on external graph data; works at "cold start".
Factual Context	Verifies claims against external knowledge graphs	HGNNR4FD [4], [14]	Unaffected by KG gaps or staleness; purely content-driven.
Multimodal Content	Fuses or aligns textual and visual information	MCOT [15], CLAAF [16], LogicDM [17]	Builds stronger text features, which can benefit any modality.
Text-Only Content	Uses deep linguistic or psychological cues	H-GIN [18, 19]	Replaces brittle fine-tuning with a robust embedding regimen.
Proposed Approach (SCU)	Synergistic SupCon + Embedding Mixup	This study	Learns a highly discriminative and generalizable text feature space.

This approach performs factual verification by validating claims within news content against an external Knowledge Graph (KG). State-of-the-art methods construct heterogeneous graphs linking claims to KG entities, enabling automated cross-document reasoning [14, 20]. For instance, HGNNR4FD [4] uses a GNN to reason over this combined graph to detect inconsistencies. While powerful, these models are bottlenecked by the completeness and timeliness of their KGs, struggling to break news or handle claims involving novel entities.

2.2. Content-Based Methods

To overcome the dependency on external signals, content-based methods analyze the intrinsic properties of the message. This research family is diverse, with recent advances pushing beyond simple text classification.

Sophisticated textual analysis models, such as H-GIN [18], construct intricate multi-channel graphs to model the syntactic, semantic, and sequential properties of text, thereby detecting the subtle linguistic patterns of propaganda. Similarly, in the related domain of citation recommendation, syntax-aware embeddings are more effective than standard representations for identifying salient sentences [21].

Others have demonstrated that a message's affective footprint is a strong signal, showing that incorporating lexicon-based emotional features can significantly improve detection performance [19, 22]. These methods confirm the long-held hypothesis that deceptive language often employs a distinct emotional palette, typically higher in negative-valence emotions, to manipulate readers [23]. Recognizing the prevalence of text-image posts, a significant research frontier has focused on multimodal detection [24]. These models, such as MCOT [15] and CLAAF [16], employ advanced techniques, including optimal transport and adaptive fusion, to align textual and visual representations.

Despite their sophistication, these content-based approaches share a common vulnerability. They typically rely on standard fine-tuning of a base model (e.g., BERT) with a cross-entropy objective. As a recent comparative study has highlighted [5], this standard training regimen can produce brittle representations that fail to distinguish between truthful and deceptive language robustly. This is particularly problematic in the context of domain adaptation, where models trained on one topic (e.g., politics) must generalize to another (e.g., health), a task where standard fine-tuning often fails [25, 26].

2.3. Representation-Driven Advances

Our work addresses this methodological gap by arguing that robust content understanding must precede the addition of external context. We draw inspiration from recent breakthroughs in representation learning. First, we adapt the principles of contrastive learning for a supervised setting using the Supervised Contrastive (SupCon) Loss [27]. This objective explicitly structures the embedding space by enforcing tight intra-class clustering and inter-class separation. Second, to improve generalization and prevent overconfidence, we incorporate Mixup [28], a powerful regularization technique applied in the embedding space. By training on convex combinations of samples, Mixup promotes smoother decision boundaries.

While these techniques exist in isolation, our work is the first to propose a synergistic framework that harmonizes them. This combination of contrastive regularization and embedding-space mixup sculpts a more resilient feature space for misinformation detection, addressing the critical need for training regimens that can withstand the evolving nature of online falsehoods [29, 30].

3. The Proposed SCU Framework

To address the challenge of learning robust and generalizable representations from text, we introduce a novel training framework designed to sculpt a highly discriminative feature space for misinformation detection. Departing from approaches that rely on external context, our method focuses on perfecting the model's fundamental understanding of the text itself.

The core of our approach is a unified training regimen that synergistically combines two key techniques: The synergy arises from a deliberate, two-stage process of regularization: (1) a supervised contrastive loss to impose a robust structure on the embedding space, and (2) an embedding-space mixup strategy for regularization and improved generalization.

As illustrated in Figure 1, these components are seamlessly integrated into a single, end-to-end trainable model built upon a standard BERT encoder.

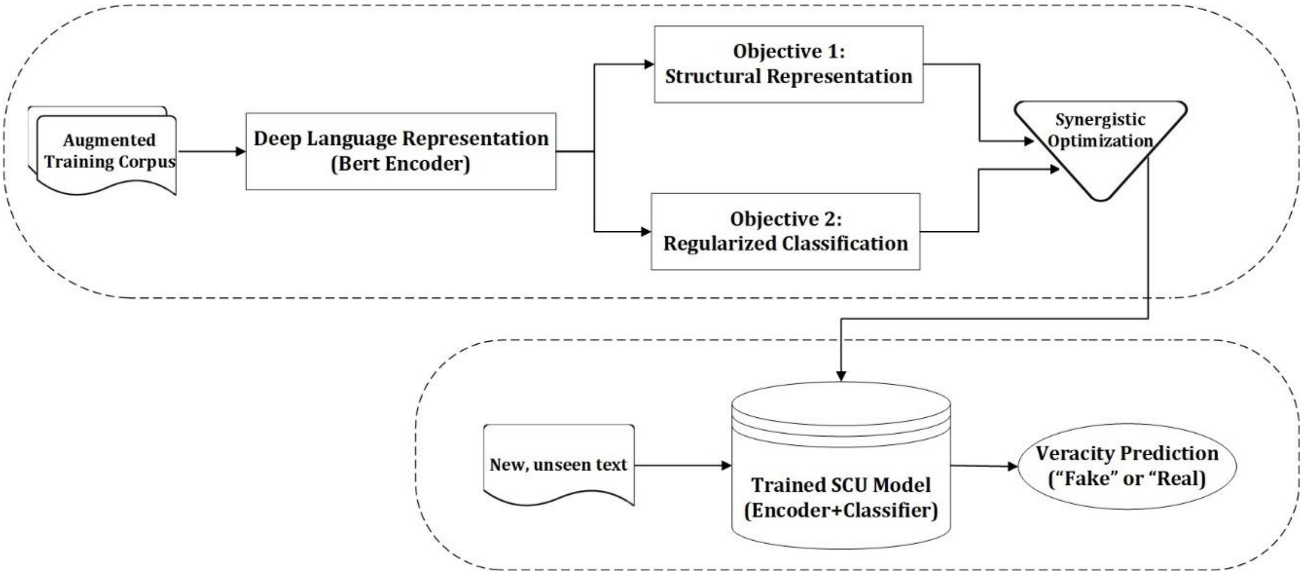
Augmented Training Corpus
Deep Language Representation (Bert Encoder)
Objective 1: Structural Representation
Objective 2: Regularized Classification
Synergistic Optimization
New, unseen text
Trained SCU Model (Encoder+Classifier)
Veracity Prediction ("Fake" or "Real)

Figure 1. The Synergistic Training Framework. BERT encodes an input text into an embedding e. This embedding drives two synergistic paths: a Representation Path, utilizing a projection head and Supervised Contrastive Loss to structure the feature space, and a Classification Path, employing Embedding-Space Mixup to regularize the final classifier. A composite loss drives end-to-end training

3.1. Architectural Overview

The framework operates on batches of labeled text samples. For each sample, the architecture bifurcates its learning objective across two parallel pathways, compelling the model to master classification and representation simultaneously.

1. **Foundational Encoding:** Each input text is first tokenized and processed by the BERT encoder. We extract the hidden state of the [CLS] token from the final layer, denoted as $e \in R^d$, which serves as a rich, context-aware embedding for the entire text. This embedding forms the foundation for the subsequent learning paths.
2. **The Representation Path:** This path transforms the general-purpose BERT embedding e into a specialized representation z that is optimized exclusively for the contrastive learning task. This is achieved by passing e through a nonlinear projection head (a two-layer MLP). The resulting vector z is then used as the sole input to the Supervised Contrastive Loss. The explicit objective of this path is to structure the latent space by maximizing intra-class similarity (pulling the embeddings of the same-class samples together) while minimizing inter-class similarity (pushing the embeddings of different-class samples apart). This process sculpts a feature space where the clusters for 'fake' and 'real' news are inherently compact and cleanly separated, providing a more robust foundation for the downstream classification path to operate upon.
3. **The Classification Path:** The original embedding **e** is fed into a linear classifier f_{cls} to produce a veracity prediction (e.g., "false" or "true"). This path is regularized using our Embedding-Space Mixup strategy and optimized via a standard Cross-Entropy loss. Its objective is to learn a highly accurate and generalizable decision boundary.

The synergy arises from this dual-path interaction: the contrastive learning on the representation path provides better-structured, more separable features. The classification path then leverages this clean feature space, allowing the Mixup regularizer to operate more effectively. Instead of interpolating between noisy or overlapping data points, Mixup creates virtual examples along the clean margin between the well-defined class clusters. This, in turn, encourages the final classifier to learn a smoother and more robust decision boundary.

3.2. The Synergistic Training Regimen

The novelty of our framework lies in its training regimen, which integrates two powerful techniques to overcome the limitations of standard fine-tuning.

Standard fine-tuning with a Cross-Entropy (CE) loss is sample-inefficient; it only considers the relationship between a single sample and its correct label, ignoring the rich information present in sample-to-sample relationships. To overcome this, we employ the Supervised Contrastive (SupCon) Loss as a powerful regularization term [27].

The intuition is to explicitly structure the embedding space by enforcing intra-class compactness and inter-class separation. To achieve this, the BERT embedding e is first mapped to a lower-dimensional representation z via a nonlinear projection head. For a minibatch of N samples, the supervised contrastive loss for a given sample (anchor) i is defined as:

$$\mathcal{L}_{SupCon}^{(i)} = -\frac{1}{|P(i)|}\sum_{p\in P(i)} log \frac{exp(sim(z_i,z_p)/\tau)}{\sum_{k=1,k\neq i}^{N} exp(sim(z_i,z_k)/\tau)} \qquad (1)$$

where $P(i) \equiv \{\mathrm{p} \mid y_p = y_i, p \neq i\}$ is the set of positive samples (i.e., other samples of the same class as i) in the batch, $sim(u, v)$ denotes cosine similarity, and $\tau \in R^+$ is a temperature scaling factor. This loss term compels the model to learn a feature space where the embeddings of all "false" news posts form a tight cluster, distinct from the "true" news cluster, making the classification task inherently easier.

Large language models are prone to overfitting and can become overconfident in their predictions. To mitigate this, we incorporate Mixup, a powerful regularization technique [28]. We apply it directly in the embedding space, which is a more stable and practical approach for NLP tasks than operating on raw text.

Mixup creates virtual training examples by forming convex combinations of existing samples. Given two [CLS] embeddings, e_i and e_j, and their corresponding one-hot labels, y_i and y_j, a new virtual sample $(\tilde{e}, \tilde{y})$ is generated as:

$$\tilde{e} = \lambda e_i + (1-\lambda)e_j \qquad (2)$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \qquad (3)$$

where $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$ for a hyperparameter α. By training the final classifier on these interpolated samples, we encourage it to learn a smoother, more linear decision boundary. This reduces the model's sensitivity to small perturbations in the input and significantly improves its generalization to unseen examples.

3.3. Final Training Objective

The framework is trained end-to-end by minimizing a composite loss function that harmonizes the goals of classification and representation learning. The final loss $\mathcal{L}_{Total}$ is a weighted sum of the standard Cross-Entropy loss on the mixed samples and the Supervised Contrastive loss on the original projected features:

$$\mathcal{L}_{Total} = (1-\beta)\cdot\mathcal{L}_{\mathrm{CE}}(f_{cls}(\tilde{e}), \tilde{y}) + \beta\cdot\mathcal{L}_{SupCon} \qquad (4)$$

where $\mathcal{L}_{CE}$ is the Cross-Entropy loss computed between the predictions of the classifier, $f_{\mathrm{cls}}(\tilde{e})$, and the interpolated labels $\tilde{y}$. The hyperparameter β, which balances the contribution of the two loss terms, is selected via a search on a held-out validation set to optimize performance. This synergistic objective ensures that the model not only learns to classify correctly but also builds a well-structured, robust, and highly discriminative feature space.

4. Experimental Setup

We conduct a rigorous evaluation to test our central hypothesis: that a content-only model trained with our synergistic framework can achieve a more robust and generalizable understanding of misinformation than models reliant on external context.

4.1. Benchmark Datasets

We evaluate our framework on two standard English-language benchmarks: Twitter15 [31] and Twitter16 [9]. These foundational rumor detection datasets contain source tweets along with their whole propagation cascades, making them ideal for a direct and fair comparison between content-only methods, such as ours, and context-aware baselines.

The key characteristics of these datasets, summarized in Table 2, make them a particularly rigorous testbed for our central hypothesis. First, their balanced class distribution ensures that performance metrics are not skewed by a majority class, providing a clean signal of a model's discriminative power. Second, their high engagement results in dense, well-defined propagation trees. This feature provides a rich source of structural information that context-aware GNN models are specifically designed to exploit, creating a challenging "home-field" advantage for our primary baselines. By benchmarking on these datasets, we are explicitly testing whether a superior content understanding can overcome the strong signals available in this abundant social context.

To ensure robust and reliable results, we employ a 5-fold stratified cross-validation protocol with a 70%/10%/20% train/validation/test split for all experiments.

Table 2. Key Statistics of Benchmark Datasets

Dataset	Total Samples	% False Class	Key Characteristics
Twitter15	742	49.9%	Balanced, High Engagement
Twitter16	412	50.2%	Balanced, High Engagement, Small-Scale

4.2. Baselines for Comparison

We benchmark our framework against a comprehensive suite of models that represent the state-of-the-art across different misinformation detection paradigms. These baselines are organized thematically to delineate their core approach and provide a multi-faceted comparison clearly.

4.2.1. Group 1: Traditional & Content-Only Baselines

These models operate on textual content without leveraging complex network structures.

- Traditional *Classifiers:* Including Decision Tree (DTC) and SVM with RBF kernel.
- *Deep Learning (Content):* BiLSTM for sequential modeling, along with powerful Transformer models (BERT, RoBERTa, XLNet, BERTweet) fine-tuned via a standard cross-entropy loss.

4.2.2. Group 2: Propagation-based Baselines

These models explicitly leverage the social network graph, focusing on the structure of how information spreads through it.

- *Hybrid GNNs:* Including a CNN-RNN hybrid (CRNN) [32] and the Bidirectional Graph Convolutional Network (Bi-GCN) [10].
- *Contrastive Graph Learning:* RDEA [33] and GACL [1], which use contrastive learning on graph structures.

4.2.3. Group 3: Hybrid and Context-Aware Baselines

This group includes sophisticated models that integrate multiple signals, representing the most powerful competitors.

- SiMiD [11]: Our primary baseline. A state-of-the-art framework that leverages user community structure and text similarity through contrastive learning.
- *Other Hybrids:* SAFE [34], which fuses text with user features; BERT-BiGRU [35]; and the meta-learning framework MetaAdapt [36].

4.3. Implementation Details

Our proposed framework and its ablations are implemented in PyTorch, building on the bert-base-uncased model from the Transformers library.

- Core Architecture: The BERT encoder produces 768-dimensional [CLS] embeddings. The non-linear projection head, used for the contrastive loss, is a two-layer MLP that maps the 768-d embedding to a 128-d space before L2 normalization.
- Training Objective: Our composite loss function (Equation 3) is optimized using the AdamW optimizer [37]. The learning rate was set to 5e-5, and the weight decay was set to 4e-4. The SupCon temperature τ was tuned to 0.03 based on validation performance. The loss-balancing hyperparameter β was set to 0.1.
- Regularization: For models using Embedding-Space Mixup, the beta distribution parameter α was set to 0.4, following established best practices.
- Ablation Models: To dissect the contribution of each component, we evaluated our full model (C+M) against an ablation using only the Contrastive loss (C).

All models were trained for a maximum of 12 epochs, using an early stopping protocol based on the weighted F1-score of the validation set, with a patience of 3 epochs to prevent overfitting and ensure a fair comparison.

4.4. Evaluation Metrics

To provide a comprehensive and robust assessment, we report on four standard classification metrics.

- Accuracy: Overall percentage of correct predictions.
- True F1 & False F1: The F1-score for each class individually, to assess performance on detecting both true and false news.
- Weighted F1-score: The F1-score averaged per class, weighted by the number of actual instances for each label. Given the balanced nature of these datasets, it serves as a primary, stable metric for comparison.

All reported results are the mean and standard deviation from a 5-fold stratified cross-validation, ensuring the statistical reliability of our findings. We use a paired t-test with a 95% confidence interval to determine if the performance differences between our model and the baselines are statistically significant.

5. Results

This section presents the empirical validation of our synergistic training framework. We designed our experiments to answer three key research questions that progressively build the case for our central hypothesis: that a sufficiently robust content-only model can surpass even those that rely on external context.

5.1. RQ1: How does our Framework Compare to Standard Content-Only Models?

Our first objective is to establish the effectiveness of our training regimen compared to standard deep learning and traditional methods that also operate exclusively on text. Table 3 presents this comparison. Our framework (C+M) significantly outperforms all other content-only baselines.

On Twitter15, our model achieves a Weighted F1-score of 0.942, surpassing the strongest Transformer baseline, RoBERTa (0.926), by +1.6 percentage points. The advantage is even more pronounced on Twitter16, where our score of 0.947 represents a +2.3 percentage point gain over the best-performing baseline, BERT (0.924). As expected, traditional classifiers like DTC and SVM-RBF are not competitive, and even a standard BiLSTM is outperformed by a large margin.

These results provide clear evidence in support of our initial premise: standard fine-tuning with a cross-entropy loss is a suboptimal strategy for this task. By synergistically combining contrastive regularization and mixup, our framework learns a fundamentally more discriminative representation from the text itself, setting a new performance standard for content-only misinformation detection.

Table 3. Performance Comparison with Content-Only Baselines. Results are reported as Mean ± Std. Dev. of Weighted F1-score

Model	Type	Twitter15	Twitter16
DTC	Traditional	0.577 ± 0.019	0.609 ± 0.063
SVM-RBF	Traditional	0.623 ± 0.043	0.621 ± 0.035
BiLSTM	Deep Learning	0.877 ± 0.062	0.700 ± 0.060
BERT	Transformer	0.918 ± 0.027	0.924 ± 0.036
RoBERTa	Transformer	0.926 ± 0.022	0.915 ± 0.028
BERTweet	Transformer	0.912 ± 0.025	0.921 ± 0.023
XLNet	Transformer	0.668 ± 0.273	0.666 ± 0.231
Our Method (SCU)	**Proposed (Content-Only)**	**0.942 ± 0.013**	**0.947 ± 0.043**

5.2. RQ2: Can our Content-Only Framework Outperform State-of-the-Art Context-Aware Models?

Having established the superiority of our framework within the content-only paradigm, we now address a more challenging question: can a model relying solely on text outperform state-of-the-art methods that leverage external social context? Table 4 presents a direct comparison against leading propagation-based and hybrid context-aware models. The results are compelling. Our purely content-based framework (C+M) achieves the highest Weighted F1-score on both datasets, surpassing all context-dependent competitors. On Twitter15, our score of 0.942 is higher than that of the top graph-based model, RDEA (0.920), and, most notably, the state-of-the-art hybrid model, SiMiD (0.931).

This trend is also observed in the Twitter16 dataset, where our model's score of 0.947 again exceeds that of SiMiD (0.937). The fact that our method, which has no access to propagation trees or user metadata, can outperform models specifically designed to exploit that information is a powerful testament to our central thesis. It suggests that the linguistic signals of misinformation, when unlocked by a robust training regimen, can be more reliable and discriminative than the structural patterns of its diffusion. This finding validates our approach and highlights the immense, often untapped, potential of advanced content analysis.

Table 4. Performance Comparison with Context-Aware State-of-the-Art. Results are reported as Mean ± Std. Dev. of Weighted F1-score

Model	Type	Twitter15	Twitter16
CRNN	Propagation-based	0.510 ± 0.020	0.537 ± 0.140
Bi-GCN	Propagation-based	0.914 ± 0.016	0.906 ± 0.044
RDEA	Propagation-based	0.920 ± 0.037	0.923 ± 0.025

GACL	Propagation-based	0.870 ± 0.032	0.914 ± 0.024
SAFE	Hybrid	0.897 ± 0.033	0.890 ± 0.027
BERT-BiGRU	Hybrid	0.812 ± 0.046	0.864 ± 0.022
MetaAdapt	Hybrid	0.639 ± 0.042	0.695 ± 0.066
SiMiD	Hybrid (SOTA)	0.931 ± 0.020	0.937 ± 0.031
Our Method (SCU)	**Proposed (Content-Only)**	**0.942 ± 0.013**	**0.947 ± 0.043**

5.3. RQ3: What is the Source of the Performance Gains? (Ablation Study)?

To validate that our framework's success stems from the proposed synergy of its components, we conducted a rigorous ablation study. We evaluated our complete model (C+M) against a version trained with only the Supervised Contrastive loss (C). The results, detailed in Table 5, reveal a clear and consistent pattern of synergistic improvement. analysis.

Table 5. Ablation Study Results (Weighted F1-score), showing the synergistic contribution of each component

Model	**Type**	**Twitter15**
Our Method (C)	0.924 ± 0.020	0.930 ± 0.021
Our Method (C+M)	0.942 ± 0.013	0.947 ± 0.043

The baseline model, which uses only Contrastive Regularization (C), already establishes a strong starting point, performing competitively with top-tier baselines. This confirms the value of explicitly structuring the feature space. However, the addition of Embedding-Space Mixup (C+M) provides a significant and consistent performance boost across both datasets. On Twitter15, the Weighted F1-score improves from 0.924 to 0.942 (+1.8 percentage points). A similar gain is observed on Twitter16, where the score rises from 0.930 to 0.947 (+1.7 percentage points).

This demonstrates that the two components are not redundant but are complementary. The contrastive loss first organizes the feature space into semantically meaningful clusters, and the mixup regularizer then ensures the final classifier learns a smoother, more generalizable decision boundary within that well-structured space. This synergy is the key driver of our framework's state-of-the-art performance.

6. Discussion

Our empirical results compel a re-examination of a widely held assumption in misinformation detection: that external social or factual context is indispensable for top-tier performance. Our work does not refute the value of context but instead argues that the quality of the core content representation is a more foundational and often overlooked element. By rigorously structuring the embedding space and regularizing the classifier, our framework consistently surpasses context-aware approaches, reframing content analysis as a first-order concern rather than a fallback option.

6.1. Implications

The primary theoretical implication of our work is the reframing of content-based methods from a mere fallback option to a foundational necessity. This principle is increasingly recognized across diverse information systems, where robust, language-independent content analysis now forms the basis for core functionalities, such as automated tag recommendation in multilingual Q&A platforms [38]. While context-aware models are undeniably powerful, our results suggest a brittle understanding of the core text can undermine their sophisticated reasoning about external signals. SCU's performance indicates that the subtle linguistic cues that differentiate truthful and deceptive language are a richer and more reliable signal than previously understood, provided they are learned through a sufficiently rigorous training regimen. This positions our framework not merely as an alternative, but as a potential "drop-in" upgrade for the content-analysis modules of future hybrid systems.

From a practical, systems-building perspective, the implications are even more direct. Our framework directly addresses the persistent “cold-start” problem that plagues real-time detection systems. Context-aware models falter on novel content that has not yet accumulated a propagation history or for which there are no up-to-date knowledge-graph entries [13]. Because our method relies solely on text, it can instantly flag new claims as they are created. This makes it ideal for applications like browser plugins or platform-integrated “pre-checks” that provide immediate, high-confidence assessments, helping to triage emerging misinformation for human moderators and protect users from the very first exposure.

6.2. Threats to Validity

While our findings are robust within our experimental design, we acknowledge several limitations that bound their generalizability and represent essential threats to validity.

A primary threat to external validity lies in the linguistic and temporal scope of our evaluation. Our experiments were conducted exclusively on English-language Twitter datasets from a specific time period (2015-2022). It is well-established that misinformation is an adversarial and rapidly evolving phenomenon. The linguistic patterns and rhetorical tactics prevalent during the events covered by our datasets may differ significantly from those used in future campaigns. For instance, the increasing sophistication of LLM-generated content presents a novel challenge that our model, trained on pre-2023 data, has not been explicitly exposed to. This highlights the critical challenge of concept drift, where the statistical properties of the target concept (i.e., "fake news") change over time.

Consequently, SCU's outstanding performance on these benchmarks does not guarantee equivalent success on other languages, platforms, or future narratives characterized by different rhetorical strategies. This underscores that no static model can be a permanent solution; any practical information system built on this framework would require periodic retraining and continuous monitoring to remain effective against new and evolving threats.

A more fundamental threat concerns the construct validity of our central claim. Our model’s architecture represents a deliberate choice to master content analysis, which necessitates a blindness to other forms of context. In cases of subtle factual errors that require external verification, knowledge-based verifiers will always excel. For uncovering coordinated inauthentic behavior or bot-driven amplification, social-context models remain indispensable. We do not claim that our method is a panacea. Instead, we assert that its robust content understanding provides a stronger foundation upon which these other contextual models can be built. Our work thus re-prioritizes, but does not eliminate, the need for a multi-faceted approach to detection.

Finally, regarding internal validity, we followed a rigorous 5-fold cross-validation protocol and used established baseline results where possible to ensure a fair comparison. The consistency of our findings across both datasets provides confidence in the reliability of our conclusions.

7. Conclusion and Future Work

In this work, we challenged a prevailing assumption in misinformation detection: that state-of-the-art performance is contingent upon access to external social or factual context. We introduced a synergistic training framework designed to sculpt a fundamentally more robust and discriminative feature space directly from textual content. By harmonizing supervised contrastive regularization with embedding-space mixup, our framework encourages a model to move beyond simple classification and learn the deep, intrinsic linguistic patterns that distinguish truthful from deceptive language. Our central finding is that the "brittleness" of standard fine-tuning has led the field to underestimate the power of intrinsic textual signals. By focusing on sculpting a superior feature space, our method provides a powerful and practical solution to the "cold-start" problem, enabling immediate, high-confidence veracity assessment.

Future work will proceed along two primary paths. First, we will explore hybrid models that fuse our advanced content encoder with the architectural strengths of context-aware systems (e.g., SiMiD, HGNNR4FD) to unlock further performance gains. This could involve adapting the framework for the real-time detection of new and emerging misinformation narratives, similar to how deep learning is applied to identify anomalous patterns in crime data [39]. Additionally, further refinement of the feature space could be achieved by integrating advanced optimization techniques, inspired by the use of meta-heuristic algorithms to improve the performance of clustering methods in other complex classification tasks [40]. Second, we aim to extend our synergistic training principles to the multimodal domain. A promising direction for a multimodal SCU (M-SCU) would be to adapt the synergy to learn a unified representation from text and images. We hypothesize this could be achieved through a three-stage process: (1) employ a joint multimodal encoder, likely based on cross-attention mechanisms, to fuse text and image features into a single embedding; (2) apply the supervised contrastive loss to this joint embedding space, which would not only separate fake and real news but also enforce semantic consistency by clustering congruent text-image pairs of the same class; and (3) perform embedding-space mixup on these unified embeddings to regularize the final classifier. Such a framework would be inherently sensitive to text-image inconsistencies while benefiting from the same robust feature space and generalizable decision boundary that make the current SCU model effective.

8. Statements & Declarations

8.1. Acknowledgments

The manuscript text has been reviewed using AI-powered proofreading tools.

8.2. Funding

The authors of this article have not received financial support from any governmental or private institutions for the research described herein.

8.3. Author Contributions

All authors contributed equally to the conceptualization, design, data analysis, and writing of the manuscript. Each author participated in the review and approval of the final manuscript.

9. References

[1] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. Science, 359(6380), 1094–1096. doi:10.1126/science.aao2998.

[2] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. doi:10.1126/science.aap9559.

[3] Guo, Y., Ji, S., Fang, X., Chiu, D. K. W., Cao, N., & Leung, H. (2025). An Unsupervised Fake News Detection

Framework Based on Structural Contrastive Learning. Cybersecurity, 8(1). doi:10.1186/s42400-024-00342-5.

[4] Xie, B., Ma, X., Wu, J., Yang, J., Xue, S., & Fan, H. (2023). Heterogeneous Graph Neural Network via Knowledge Relations for Fake News Detection. 35th International Conference on Scientific and Statistical Database Management, 1–11. doi:10.1145/3603719.3603736.

[5] Kuntur, S., Krzywda, M., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study. Electronics (Switzerland), 13(23), 4784. doi:10.3390/electronics13234784.

[6] Mridha, M. F., Keya, A. J., Hamid, Md. A., Monowar, M. M., & Rahman, Md. S. (2021). A Comprehensive Review on Fake News Detection With Deep Learning. IEEE Access, 9, 156151–156170. doi:10.1109/access.2021.3129329.

[7] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining, 13(1), 1–36. doi:10.1007/s13278-023-01028-5.

[8] Farhangian, F., Cruz, R. M. O., & Cavalcanti, G. D. C. (2024). Fake news detection: Taxonomy and comparative study. Information Fusion, 103, 102140. doi:10.1016/j.inffus.2023.102140.

[9] Ma, J., Gao, W., & Wong, K. F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 1, 1980–1989. doi:10.18653/v1/p18-1184.

[10] Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 34(1), 549–556. doi:10.1609/aaai.v34i01.5393.

[11] Ozcelik, O., Toraman, C., & Can, F. (2025). Detecting Misinformation on Social Media using Community Insights and Contrastive Learning. ACM Transactions on Intelligent Systems and Technology, 16(2). doi:10.1145/3709009.

[12] Truică, C. O., Apostol, E. S., & Karras, P. (2024). DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection. Knowledge-Based Systems, 294, 111715. doi:10.1016/j.knosys.2024.111715.

[13] Jovanović, A., & Ross, B. (2023). Rumour Detection in the Wild: A Browser Extension for Twitter. 3rd Workshop for Natural Language Processing Open Source Software, NLP-OSS 2023, Proceedings of the Workshop, 130–140. doi:10.18653/v1/2023.nlposs-1.15.

[14] Koloski, B., Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. Neurocomputing, 496, 208–226. doi:10.1016/j.neucom.2022.01.096.

[15] Shen, X., Huang, M., Hu, Z., Cai, S., & Zhou, T. (2024). Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. Frontiers in Computer Science, 6. doi:10.3389/fcomp.2024.1473457.

[16] Mu, G., Chen, C., Li, X., Chen, Y., Dai, J., & Li, J. (2025). CLAAF: Multimodal fake information detection based on contrastive learning and adaptive Agg-modality fusion. PLoS ONE, 20(5 MAY), 322556. doi:10.1371/journal.pone.0322556.

[17] Liu, H., Wang, W., & Li, H. (2023). Interpretable Multimodal Misinformation Detection with Logic Reasoning. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 9781–9796. doi:10.18653/v1/2023.findings-acl.620.

[18] Ahmad, P. N., Guo, J., AboElenein, N. M., Haq, Q. M. ul, Ahmad, S., Algarni, A. D., & A. Ateya, A. (2025). Hierarchical graph-based integration network for propaganda detection in textual news articles on social media. Scientific Reports, 15(1). doi:10.1038/s41598-024-74126-9.

[19] Farhoudinia, B., Ozturkcan, S., & Kasap, N. (2024). Emotions unveiled: detecting COVID-19 fake news on social media. Humanities and Social Sciences Communications, 11(1), 504. doi:10.1057/s41599-024-03083-5.

[20] Wu, X., Huang, K.-H., Fung, Y., & Ji, H. (2022). Cross-document Misinformation Detection based on Event Graph Reasoning. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. doi:10.18653/v1/2022.naacl-main.40.

[21] Roostaee, M. (2022). Citation Worthiness Identification for Fine-Grained Citation Recommendation Systems. Iranian Journal of Science and Technology - Transactions of Electrical Engineering, 46(2), 353–365. doi:10.1007/s40998-021-00472-3.

[22] Giachanou, A., Ghanem, B., Ríssola, E. A., Rosso, P., Crestani, F., & Oberski, D. (2022). The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. Data & Knowledge Engineering, 138, 101960. doi:10.1016/j.datak.2021.101960.

[23] Pillai, S. E. V. S., & Hu, W.-C. (2023). Misinformation Detection Using an Ensemble Method with Emphasis on Sentiment and Emotional Analyses. 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA), 295–300. doi:10.1109/sera57763.2023.10197706.

[24] Abdali, S., Shaham, S., & Krishnamachari, B. (2024). Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. ACM Computing Surveys, 57(3). doi:10.1145/3697349.

[25] Yue, Z., Zeng, H., Kou, Z., Shang, L., & Wang, D. (2022). Contrastive Domain Adaptation for Early Misinformation Detection. Proceedings of the 31st ACM International

Conference on Information & Knowledge Management, 2423–2433. doi:10.1145/3511808.3557263.

[26] Nan, Q., Cao, J., Zhu, Y., Wang, Y., & Li, J. (2021). MDFEND: Multi-domain Fake News Detection. Proceedings of the 30th ACM International Conference on Information &amp; Knowledge Management, 3343–3347. doi:10.1145/3459637.3482139.

[27] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. 34rd Conference on Neural Information Processing Systems (NeurIPS 2020), 33, 18661-18673, 6-12, December, 2020, Vancouver, Canada.

[28] Huang, L., Zhang, C., & Zhang, H. (2020). Self-adaptive training: beyond empirical risk minimization. Advances in neural information processing systems, 33, 19365-19376.

[29] Cavus, N., Goksu, M., & Oktekin, B. (2024). Real-time fake news detection in online social networks: FANDC Cloud-based system. Scientific Reports, 14(1). doi:10.1038/s41598-024-76102-9.

[30] Wan, H., Feng, S., Tan, Z., Wang, H., Tsvetkov, Y., & Luo, M. (2024). DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. Findings of the Association for Computational Linguistics ACL 2024, 2637–2667. doi:10.18653/v1/2024.findings-acl.155.

[31] Ma, J., Gao, W., & Wong, K.-F. (2017). Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. doi:10.18653/v1/p17-1066.

[32] Liu, Y., & Wu, Y. F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 32(1), 354–361. doi:10.1609/aaai.v32i1.11268.

[33] He, Z., Li, C., Zhou, F., & Yang, Y. (2021). Rumor Detection on Social Media with Event Augmentations. SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020–2024. doi:10.1145/3404835.3463001.

[34] Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-Aware Multi-modal Fake News Detection. Advances in Knowledge Discovery and Data Mining. PAKDD 2020. Lecture Notes in Computer Science, vol 12085. Springer, Cham, Switzerland. doi:10.1007/978-3-030-47436-2_27.

[35] Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. Knowledge-Based Systems, 274, 110642. doi:10.1016/j.knosys.2023.110642.

[36] Yue, Z., Zeng, H., Zhang, Y., Shang, L., & Wang, D. (2023). MetaAdapt: Domain Adaptive Few-Shot Misinformation Detection via Meta Learning. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.286.

[37] D'Angelo, F., Andriushchenko, M., Varre, A. V., & Flammarion, N. (2024). Why do we need weight decay in modern deep learning?. Advances in Neural Information Processing Systems, 37, 23191-23223.

[38] Roostaee, M. (2024). Language-independent Profile-based Tag Recommendation for Community Question Answering Systems. International Journal of Engineering Transactions C: Aspects, 37(12), 2547–2559. doi:10.5829/ije.2024.37.12c.13.

[39] Dorrani, Z. (2025). Anomaly Detection in Emerging Crimes with Deep Autoencoder Architecture. Contributions of Science and Technology for Engineering, 2(3), 45-56. doi:10.22080/cste.2025.28900.1023.

[40] Rasouli, A., Mikaeil, R., Atalou, S., & Esmaeilzadeh, A. (2025). Optimizing Traditional Clustering Methods Using Meta-Heuristic Algorithms for Joint Set Identification in Copper Mines. Contributions of Science and Technology for Engineering, 2(3), 57-72. doi:10.22080/cste.2025.29024.1032.