SemIRNet: A Semantic Irony Recognition Network for Multimodal Sarcasm Detection

1st Jingxuan Zhou †
University of New South Wales
Canberra, Australia
zhou20040626@outlook.com

2nd Yuehao Wu † *
University of Sydney
Sydney, Australia
yuwu6640@uni.sydney.edu.au

3rd Yibo Zhang
University of New South Wales
Canberra, Australia
bernie.zhangyibo@gmail.com

4th Yeyubei Zhang
University of Pennsylvania
Philadelphia, United States
joycezh@alumni.upenn.edu

5th Yunchong Liu
University of Pennsylvania
Philadelphia, United States
yunchong@alumni.upenn.edu

6th Bolin Huang
University of Southern California
California, United States
bolinhua@usc.edu

7th Chunhong Yuan
Kazan Federal University
Kazan, Russia
ChYuan@kpfu.ru

2025 10th International Conference on Information and Network Technologies (ICINT) | 979-8-3315-3675-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICINT65528.2025.11030873

***Abstract*—Aiming at the problem of difficulty in accurately identifying graphical implicit correlations in multimodal irony detection tasks, this paper proposes a Semantic Irony Recognition Network (SemIRNet). The model contains three main innovations: (1) The ConceptNet knowledge base is introduced for the first time to acquire conceptual knowledge, which enhances the model"s common-sense reasoning ability; (2) Two cross-modal semantic similarity detection modules at the word level and sample level are designed to model graphic-textual correlations at different granularities; and (3) A contrastive learning loss function is introduced to optimize the spatial distribution of the sample features, which improves the separability of positive and negative samples. Experiments on a publicly available multimodal irony detection benchmark dataset show that the accuracy and F1 value of this model are improved by 1.64% and 2.88% to 88.87% and 86.33%, respectively, compared with the existing optimal methods. Further ablation experiments verify the important role of knowledge fusion and semantic similarity detection in improving the model performance.**

***Index Terms*—Multimodel Learning, Knowledge Fusion, Semantic Similarity, Comparative Learning**

I. Introduction

With the rapid development of social media, users generate a large number of multimodal messages containing images and texts. In these messages, ironic expressions are becoming more and more common as a specific linguistic phenomenon. Irony is defined as "saying or writing the opposite of what is actually intended, or speaking with the intention of making others feel stupid or letting them know that you are angry". Irony detection is crucial to the task of sentiment analysis because ironic language often expresses the opposite emotional polarity of its intended meaning compared to its literal meaning [1]–[4].

Early irony language detection mainly utilized textual information. With the increase of multimodal data, more and more studies have started to focus on multimodal irony detection. Existing multimodal irony detection models are mainly based on two approaches: (1) Attention mechanism-based models capture inter-modal inconsistencies by designing different deformation structures of the attention mechanism; and (2) graph neural network-based models establish correlations of multimodal information by constructing cross-modal graph networks. However, these approaches mainly focus on the surface features of images and texts, ignoring the importance of commonsense knowledge for detecting and understanding non-literal expressions. In the field of affective computing, the integration of commonsense knowledge has been shown to help improve the performance of models, as acquiring commonsense knowledge through task-specific dataset learning alone is inherently difficult [5]–[7].

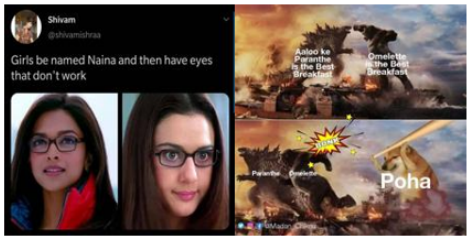

Fig. 1. Examples of non-ironic and ironic samples.

Based on this, this paper constructs a knowledge fusion model that conforms to the human cognitive style from the perspective of human recognition of ironic language. As shown in Fig. 1. in non-ironic multimodal information, the semantic information in text and images is often explicitly related. On the contrary, in ironic samples, the semantic information in text and images tends to be opposite or implicit, i.e., the image-text information is implicitly correlated. Such implicitly correlated multimodal data need to be identified with some commonsense information, as implicit emotional expressions often require a more conceptual understanding of words in different situations.

(1) The main contributions of this paper are as follows: A novel Semantic Irony Recognition Network that integrates the ConceptNet knowledge base for multimodal irony detection

†These authors contributed equally to this work.

is proposed. To the best of our knowledge, this is the first model to incorporate conceptual knowledge to enhance irony detection accuracy.

(2) A new multimodal information processing method is designed, featuring word-level and sample-level cross-modal semantic similarity detection modules. These modules assess the semantic consistency of different modalities, enabling the model to extract features relevant to irony detection.

(3) Contrastive learning is introduced to distinguish ironic (positive) and non-ironic (negative) samples. A contrastive learning loss function is employed to refine multimodal feature representation. The proposed model achieves state-of-the-art performance on public benchmark datasets.

The paper is organized as follows: section 2 introduces the related work; section 3 describes the proposed SemIRNet model in detail; section 4 gives the experimental results and analysis; and section 5 concludes the paper.

II. Relater work

This section presents related work in two main areas: multimodal learning and multimodal irony detection.

A. Multimodal Learning

Multimodal learning is a method of content analysis and understanding using multiple modes of information delivery. Since multimedia data usually contains multiple modal information such as image and text, multimodal learning has become the main method for multimedia content analysis. Some representative works in image-text information fusion have emerged in recent years: (1) The ERNIE-ViLM model utilizes structured knowledge in scene graphs to enable the model to perform fine-grained semantic alignment. (2) The VIVO model uses Image-Tag for pre-training to align semantic labels with region features in images. (3) The RpBERT model uses multimodal Bidirectional Encoder Representations from Transformers (BERT) for entity recognition tasks, and the proposed relation propagation mechanism allows for better utilization of visual information based on the correlation between images and text.

B. Multimodal Irony Detection

1) Text-based Approach: Early research on text-based irony detection primarily relied on traditional natural language processing techniques, such as lexical annotation, syntactic analysis, and linguistic feature extraction, to identify ironic expressions. With advancements in deep learning, researchers introduced neural network models to enhance detection performance. For instance, textCNN [8] employs convolutional neural networks to capture local semantic features of text. The emergence of pre-trained language models further revolutionized irony detection by leveraging large-scale pre-training to achieve robust language understanding. These models also introduced parameter-sharing techniques, reducing model size while maintaining performance, thereby significantly improving text-based irony detection accuracy. However, relying solely on textual features presents limitations. These methods struggle with ironic expressions that require background knowledge, fail to incorporate auxiliary information such as images in social media, and overly depend on specific linguistic patterns, limiting their generalization. These challenges underscore the necessity of developing multimodal irony detection methods.

2) Text-based Approach: With the rapid growth of social media, user-generated content increasingly exhibits multimodal characteristics, combining text and images [9]. Compared to text-only methods, multimodal irony detection emphasizes inter-modal information fusion and interaction to better capture users' expressive intent. Two primary technical approaches have emerged: attention-based methods and graph neural networks (GNNs).

(1) Attention-based methods dynamically establish associations between modalities to capture irony-related features. For example, Pan et al. proposed a cross-modal attention model based on BERT [10], which detects intra-modal and inter-modal semantic inconsistencies. Additionally, researchers have developed end-to-end text-visual fusion models based on Transformers, further improving modality integration.

(2) GNN-based methods construct cross-modal interaction graphs to model ironic features. This graph-based approach is particularly effective in handling the complex relationships between multimodal data [11].

Despite notable advancements, existing multimodal irony detection approaches face key challenges. They often prioritize surface feature alignment over deep semantic understanding and fail to incorporate common-sense knowledge, leading to poor performance in complex scenarios [12]. Additionally, simplistic modality fusion strategies fail to fully leverage the complementary strengths of different modalities [13]. These limitations highlight the need for cognitively inspired multimodal irony detection models that better align with human understanding.

III. Method

In this paper, we propose a Semantic Irony Recognition Network called SemIRNet for multimodal irony detection. The model architecture contains the following main components:

(1) Text and image feature extraction module: use pre-trained BERT and ResNet to encode text and image information respectively

(2) ConceptNet-based Knowledge Enhancement Module: Introducing Conceptual Knowledge to Enhance Common Sense Reasoning in Models

(3) Cross-modal semantic similarity detection module: two levels: word-level and sample-level

(4) Comparative learning optimization module: improving feature space distribution

A. Technical Details

For text input, the model utilizes a pre-trained BERT encoder, which separately processes the main text and title text, outputting corresponding feature vectors. For image data,

visual features are extracted using ResNet, and an average pooling layer is applied to obtain a fixed-dimensional feature representation.

To enhance semantic understanding, we incorporate the ConceptNet knowledge base, enriching text and image attributes at the conceptual level [14]. Specifically, ConceptNet is queried to retrieve related concepts and relationships associated with text and image attribute words, expanding semantic information [15]. This conceptual information is then encoded into vector form for subsequent semantic similarity computation.

For knowledge enhancement, while simple feature alignment effectively captures explicit modal associations, our framework integrates deeper semantic understanding through ConceptNet. For example, when processing contrasting pairs like "sunny photo" and "bad day", the knowledge enhancement module establishes semantic connections beyond surface-level feature matching by leveraging conceptual relationships from ConceptNet. This approach balances computational efficiency with semantic comprehension [16]. The knowledge integration process involves querying related concepts, encoding them into vector form, and computing semantic similarities.

To achieve multi-level semantic similarity detection, we design two mechanisms:

(1) Word-level similarity detection: Conceptual representations of text and image attributes are compared using matrix operations, with maximum pooling extracting the most significant semantic associations. This allows the model to capture fine-grained cross-modal correspondences.

(2) Sample-level spatial mapping: A spatial mapping mechanism aligns feature spaces across modalities. Text features (including main text and captions) are concatenated with image features, and a sample covariance matrix is computed to derive a mapping matrix. This transformation projects features into a shared semantic space, facilitating similarity computation.

To further enhance classification performance, we introduce a triad-based contrastive learning mechanism. During training, for each anchor sample, positive samples (from the same category) and negative samples (from different categories) are selected. The model optimizes feature space distribution by minimizing the distance between anchor and positive samples while maximizing the distance between anchor and negative samples.

During training, the model is trained end-to-end using the Adam optimizer, with a learning rate of 1×10^{-5} and a batch size of 32. In the inference stage, input data undergo feature extraction, knowledge enhancement, and semantic similarity detection, ultimately producing the irony detection result.

B. Implementation Details

This section details the key settings of the model implementation. Table I summarizes the main model configuration parameters:

In the process of model training, we adopt a series of strategies to ensure the stability of performance and generalization ability. Firstly, in the data preprocessing stage, the text data are randomly masked with the mask ratio set to 15%, and the image data are enhanced with random cropping, horizontal flipping and normalization. These data enhancement techniques effectively improve the robustness of the model.

TABLE I
KEY CONFIGURATION PARAMETERS OF SEMIRNET.

Module	Parameter	Setting
Text Encoding	Pre-trained Model	BERT-base
	Hidden Dimension	768
	Max Length	128 token
Image Encoding	Backbone	ResNet-152
	Input Size	224×224
	Caption Model	MobileNetV3
Knowledge Integration	ConceptNet Dimension	300
Training	Optimizer	Adam
	Learning Rate	1e−5
	Batch Size	32
	Contrastive Margin	0.5
	Loss Weight λ	0.1

The training adopts a staged strategy: firstly, the text and image encoders are pre-trained to obtain the basic feature representation capability, then the ConceptNet knowledge vectors are loaded for knowledge enhancement, and finally, the end-to-end model training is carried out. During the training process, we use the performance of the validation set as the early stopping criterion, and dynamically adjust the hyperparameters accordingly.

The whole training process is carried out on a single RTX 4060 graphics card. Through the optimization of the above implementation details, the model is able to achieve stable and excellent performance. The experimental results show that the choice of these parameter configurations and training strategies is crucial to realize the potential of the model in the multimodal irony detection task.

IV. EXPERIMENTS

A. Main Results

We evaluated the performance of SemIRNet on a publicly available multimodal irony detection benchmark dataset. Fig. 2. and Table II show the experimental results:

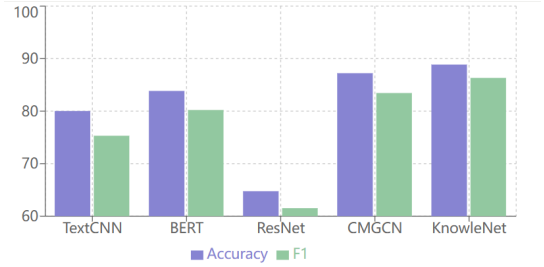

Fig. 2. Visualization of model performance comparison on Dataset-1.

The following key findings can be observed from the experimental results: Among the unimodal methods, the text-based methods generally outperform the image-based methods. Among them, the BERT model shows the best unimodal

TABLE II
PERFORMANCE COMPARISON ON DATASET-1 (%).

Model	Accuracy	Precision	Recall	F1-score
TextCNN	80.03	74.29	76.39	75.32
BERT	83.85	78.72	82.27	80.22
ResNet	64.76	54.41	70.80	61.53
ViT	67.83	57.93	70.07	63.43
HFM	83.44	76.57	84.15	80.18
D&R Net	84.02	77.97	83.42	80.60
CMGCN	87.23	-	-	83.45
SemIRNet	**88.87**	**88.59**	**84.18**	**86.33**

TABLE III
ABLATION STUDY RESULTS ON DATASET-1 (%).

Model Variant	Accuracy	ΔAcc	F1-score	ΔF1	Macro -F1	ΔMacro -F1
SemIRNet	**88.87**	-	**86.33**	-	**88.51**	-
w/o Knowledge	85.89	-2.98	82.21	-4.12	84.58	-3.93
w/o Semantic	85.12	-3.75	81.85	-4.48	83.38	-5.13
w/o Contrastive	86.46	-2.41	83.32	-3.01	85.63	-2.88

performance, reaching 83.85% accuracy and 80.22% F1 value. This indicates that textual information is more discriminative than visual information in irony detection tasks. Multimodal methods have significantly improved their performance by fusing text and image information. Earlier multimodal methods such as Hierarchical Fusion Model(HFM) and Detection and Recognition Network(D&R) Net have demonstrated better performance than unimodal methods. The latest CMGCN model based on graph neural network further improves the accuracy to 87.23%. Our proposed SemIRNet model achieves optimal performance on all evaluation metrics by introducing knowledge enhancement and multi-level semantic similarity detection. Specifically:

(1) Accuracy of 88.87 per cent, an improvement of 1.64 per cent compared to CMGCN

(2) F1 value reached 86.33%, an improvement of 2.88% compared to CMGCN

(3) Achieved a balanced improvement in both precision and recall rates

These results confirm the effectiveness of knowledge fusion and semantic alignment in improving the performance of multimodal irony detection, especially when dealing with complex samples that require deep semantic understanding, our approach shows significant advantages.

B. Ablation Study

In order to verify the effectiveness of each key component of the model, we conducted detailed ablation experiments. Fig. 3. and Table III demonstrate the experimental results:

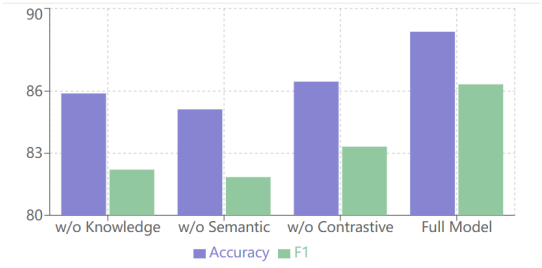

Fig. 3. Visualization of ablation experiment results.

(1) Knowledge Enhancement Module: Removing the Knowledge Enhancement Module (w/o Knowledge) causes a 2.98% decrease in accuracy, and a more significant decrease in F1 value (4.12%), which suggests that knowledge fusion is crucial to the model"s discriminative ability, which validates the need to introduce ConceptNet for conceptual-level semantic enhancement.

(2) Semantic Similarity Detection: The removal of Semantic Similarity Detection (w/o Semantic) causes the largest performance degradation (-3.75% accuracy), and Macro-F1 decreases by 5.13%, which indicates that this module is particularly important for dealing with data imbalance. The results confirm the key role of multilevel Semantic Similarity Detection in capturing inter-modal relationships.

(3) Contrastive Learning Optimization: Removing the contrastive learning loss (w/o Contrastive) decreases the accuracy by 2.41%. The performance degradation is relatively small but still significant, indicating that contrastive learning does help the model to learn a more discriminative feature representation.

These experimental results clearly show that each component of the model contributes significantly to the final performance. In particular, the importance of the semantic similarity detection module is most prominent, which is in line with our original design intention of emphasizing deep semantic understanding. Also, the synergistic effect of knowledge enhancement and comparison learning is shown to be a key factor in improving the model's performance.

The experimental results demonstrate that our modal fusion strategy effectively balances performance and computational complexity. Ablation studies validate the contribution of each component and identify opportunities for future improvement. Looking ahead, more sophisticated fusion approaches could be explored to further enhance the model's capabilities. Potential extensions include adaptive attention mechanisms that dynamically adjust based on modal consistency [17], multi-level semantic alignment frameworks [18], and cognitive-inspired information integration patterns [19]. These advanced techniques could improve the model's ability to interpret complex ironic expressions while maintaining computational efficiency. The performance gains observed in our experiments suggest that such enhancements could lead to meaningful improvements in multimodal irony detection.

C. Qualitative Analysis

(1)Performance Advantage: SemIRNet achieves optimal performance on several datasets, especially on Dataset-1 where the accuracy and F1 value are improved by 1.64% and 2.88%, respectively, compared to the best baseline CMGCN. This

confirms the effectiveness of our proposed knowledge fusion-based approach on the multimodal irony detection task.

(2)Module contribution: The ablation experiments clearly demonstrate the importance of each technology module: the semantic similarity detection module contributes the most (ΔAcc: -3.75%), the knowledge enhancement module is the second most important (ΔAcc: -2.98%), and the contrastive learning optimisation also plays a significant role (Δ Acc: -2.41%).

(3)Methodological innovations: The experimental results validate our three main innovations: the introduction of ConceptNet for knowledge enhancement indeed improves the semantic comprehension of the model, multi-level semantic similarity detection effectively captures the complex relationships between modalities, and contrastive learning optimization improves the distribution structure of the feature space.

While our experimental results show that surface feature matching can achieve satisfactory accuracy (86.33% F1-score) with relatively low computational complexity, more sophisticated fusion schemes can be explored to further enhance the model's performance. Advanced approaches such as hierarchical attention fusion [20], dynamic modal alignment [21], and cognitive-inspired integration mechanisms [22] could potentially improve the model's ability to capture implicit semantic relationships.

V. Conclusion

In this paper, we propose Semantic Irony Recognition Network (SemIRNet) to address the challenge of accurately identifying implicit associations in multimodal irony detection. Experiments on multiple public datasets demonstrate that our model improves accuracy and F1-score by 1.64% and 2.88%, respectively, compared to existing state-of-the-art methods. Ablation studies further validate the effectiveness of each module, particularly highlighting the significant contributions of semantic similarity detection and knowledge enhancement to overall performance.

For future work, we aim to explore several promising directions to further enhance the model's capabilities. First, we plan to investigate advanced fusion mechanisms, such as hierarchical attention networks [1] and dynamic modal alignment [2], to improve the model's ability to capture implicit semantic relationships between visual and textual content. Additionally, we intend to develop lightweight yet effective knowledge enhancement techniques [6] that can enrich semantic understanding while maintaining computational efficiency.

Another key direction is the investigation of cross-domain adaptation methods, which could significantly improve the model's generalization across different social media platforms and content types. These extensions are expected to lead to more robust and efficient multimodal irony detection systems, addressing current limitations in handling complex ironic expressions and context-dependent cases, while ensuring practical computational efficiency for real-world applications.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," *arXiv preprint arXiv:1606.07356*, 2016.

[3] C. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, pp. 155–177, 2015.

[4] C. Finn, "Learning to learn with gradients," Master's thesis, University of California, Berkeley, 2018.

[5] R. Anderson, B. Stenger, V. Wan *et al.*, "Expressive visual text-to-speech using active appearance models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 3382–3389.

[6] J. Andreas, M. Rohrbach, T. Darrell *et al.*, "Neural module networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 39–48.

[7] G. Andrew, R. Arora, J. Bilmes *et al.*, "Deep canonical correlation analysis," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1247–1255.

[8] A. Rakhlin, "Convolutional neural networks for sentence classification," *GitHub*, vol. 6, p. 25, 2016.

[9] L. Li, R. Wang, M. Zou, F. Guo, and Y. Ren, "Enhanced resnet-50 for garbage classification: Feature fusion and depth-separable convolutions," *PloS one*, vol. 20, no. 1, p. e0317999, 2025.

[10] M. Pan, J. Wang, J.-X. Huang *et al.*, "A probabilistic framework for integrating sentence-level semantics via bert into pseudo-relevance feedback," *Information Processing & Management*, vol. 59, no. 1, p. 102734, 2022.

[11] G. Zhao, P. Li, Z. Zhang, F. Guo, X. Huang, W. Xu, J. Wang, and J. Chen, "Towards sar automatic target recognition: Multi-category sar image classification based on light weight vision transformer," in *2024 21st Annual International Conference on Privacy, Security and Trust (PST)*. IEEE, 2024, pp. 1–6.

[12] Y. Cao, J. Dai, Z. Wang, Y. Zhang, X. Shen, Y. Liu, and Y. Tian, "Systematic review: Text processing algorithms in machine learning and deep learning for mental health detection on social media," *arXiv preprint arXiv:2410.16204*, 2024.

[13] L. Li, Z. Li, F. Guo, H. Yang, J. Wei, and Z. Yang, "Prototype comparison convolutional networks for one-shot segmentation," *IEEE Access*, 2024.

[14] H.-C. Dan, Z. Huang, B. Lu, and M. Li, "Image-driven prediction system: Automatic extraction of aggregate gradation of pavement core samples integrating deep learning and interactive image processing framework," *Construction and Building Materials*, vol. 453, p. 139056, 2024.

[15] Y. Wei, D. Zhang, M. Gao, Y. Tian, Y. He, B. Huang, and C. Zheng, "Breast cancer prediction based on machine learning," *Journal of Software Engineering and Applications*, vol. 16, pp. 348–360, 2023.

[16] D. Zhang, F. Zhou, Y. Wei, X. Yang, and Y. Gu, "Unleashing the power of self-supervised image denoising: A comprehensive review," *arXiv preprint arXiv:2308.00247*, 2023.

[17] R. Panda, C. Chen, Q. Fan *et al.*, "Adamml: Adaptive multi-modal learning for efficient video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7576–7585.

[18] Y. Xu, Y. Li, M. Xu *et al.*, "Hka: A hierarchical knowledge alignment framework for multimodal knowledge graph completion," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 8, pp. 1–19, 2024.

[19] X. Chen, H. Xie, G. Cheng *et al.*, "A decade of sentic computing: topic modeling and bibliometric analysis," *Cognitive Computation*, vol. 14, no. 1, pp. 24–47, 2022.

[20] H. Tao and Q. Duan, "Hierarchical attention network with progressive feature fusion for facial expression recognition," *Neural Networks*, vol. 170, pp. 337–348, 2024.

[21] A. Nadeem, A. Hilton, R. Dawes *et al.*, "Cad-contextual multi-modal alignment for dynamic avqa," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7251–7263.

[22] X. Chen, H. Xie, S. Qin *et al.*, "Cognitive-inspired deep learning models for aspect-based sentiment analysis: A retrospective overview and bibliometric analysis," *Cognitive Computation*, pp. 1–39, 2024.