Computers and Artificial Intelligence；Vol.3, No.4; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n4.2934
Published by: Zeus Press

A Review of Deep Sarcasm Detection Research Based on Incongruity Theory

Minkai Huang*
School of Statistics and Data Science, Nankai University, Tianjin, China
**Corresponding author: Minkai Huang.*

Abstract

As one of the most classic types of counterfactual discourse, the core feature of satire is “subtext”, which involves detecting the cognitive conflict of this “hidden meaning within the words”. This paper uses incongruity as a clue to review the related work on deep sarcasm detection and provides a corresponding classification summary. Specifically, after briefly introducing the psychological definition of sarcasm, this paper analyzes the three research perspectives (semantic perspective, emotional perspective, and factual conflict) currently used for modeling sarcasm computation, and further presents the challenges faced by the current methods. Finally, it summarizes how DNN models such as Transformer and graph networks implicitly or explicitly measure the inconsistency between these modalities; it analyzes the existing problems of poor interpretability, insufficient modality alignment, and lack of common sense integration in current models, and proposes the future development trend of integrating multimodal large models with cognitive science research.

Keywords

sarcasm detection, incongruity theory, multimodal learning, deep learning, natural language processing

1. Introduction

As a common way of expressing thoughts and emotions, sarcasm is characterized by a mismatch between the literal meaning of an expression and the speaker’s intended emotion [1]. In recent years, with the rapid growth of user-generated content on social media, sarcasm detection has received increasing attention in natural language processing (NLP) and sentiment analysis [2]. Correctly identifying sarcasm helps distinguish real emotions from sarcastic expressions, which is useful in many applications, such as public opinion analysis and political opinion monitoring [3].

Early research mainly studied sarcasm detection using text only. Most methods focused on linguistic features, contextual information, and lexical resources to recognize sarcastic intentions [1, 4]. However, as images have become an important part of social media communication, many users now express sarcasm through both text and images [5]. Because of this change, research has gradually moved from text-based sarcasm detection to multimodal sarcasm detection [6].

Compared with many other multimodal tasks, multimodal sarcasm is usually caused by inconsistency rather than agreement between different modalities [3]. In many cases, the text may express a positive attitude, while the accompanying image conveys negative information, and this mismatch creates a sarcastic meaning [7]. Therefore, recent studies have paid more attention to modeling the differences between visual and textual information.

One line of research applies graph neural networks (GNNs) to learn the relationships within and across different modalities. Liang et al. [8] built interactive intra-modal and cross-modal structures for sarcasm detection. Later, Liang et al. [9] proposed a cross-modal graph convolutional network (CMGCN), which explicitly models the incongruity between visual objects and textual phrases. Liu et al. [10] further introduced a hierarchical framework with knowledge enhancement. Ke et al. [11] proposed a semantic–emotional incongruity method and adopted contrastive learning to capture more fine-grained conflicts between modalities.

Another issue in multimodal sarcasm detection is that many sarcastic expressions depend on background knowledge and common sense. Yue et al. [12] argued that previous studies did not make sufficient use of external knowledge. They therefore proposed a knowledge-based model that introduces ConceptNet to measure cross-modal semantic similarity, which helps identify weak semantic correlations that are common in sarcastic expressions [12].

Although many methods have been proposed, multimodal sarcasm detection is still facing several challenges. This paper reviews recent studies in this area and summarizes existing methods according to their main ideas and modeling strategies.

2. Incongruity Theory of Sarcasm and Its Computational Modeling

The incongruity theory suggests that sarcasm comes from a conflict between what people expect and what they actually receive. In multimodal sarcasm detection, this idea is usually represented in a multimodal feature space and modeled from the following three aspects:

2.1 Modeling Incongruity at the Factual Level

Factual-level incongruity refers to the mismatch between the semantic information contained in the text and the image. In most studies, this mismatch is represented by the distance between image features and text features in a shared embedding space. If the text and the image describe similar facts, their feature representations are usually close to each other. By contrast, if they convey conflicting or unrelated facts, the similarity becomes lower, making sarcastic expressions easier to identify.

Early studies mainly measured this inconsistency using simple metrics, such as Euclidean distance or cosine similarity. More recent work often relies on pre-trained vision-language models, such as CLIP, to obtain richer semantic representations. The similarity between the two modalities is then calculated to capture factual conflicts at a more global level.

2.2 Modeling Incongruity at the Emotional Level

Emotional incongruity is one of the most common signals used for multimodal sarcasm detection. In many sarcastic posts, the emotions expressed by the text and the image are not consistent. A common approach is to analyze the sentiment of each modality separately. For images, CNN-based models are often used to identify positive or negative emotions, while the sentiment of the text is usually obtained with sentiment lexicons or sentiment classifiers. The results are then compared to determine whether emotional inconsistency exists.

When the sentiment expressed in the text is opposite to that of the image, such as positive text together with a negative image, the sample is more likely to be sarcastic. Some studies further model this difference with graph-based methods or introduce sentiment knowledge bases, such as SenticNet, to capture more detailed emotional inconsistencies.

2.3 Dual Perception and Interaction Mechanisms

An effective sarcasm detector usually requires the incongruity between the factual aspect and the emotional aspect. “Factual incongruity” images might be wrongly labeled, while “emotional incongruity” might be without semantic basis. Therefore, recent research methods attempt to use two perception approaches to capture sarcastic phenomena: Firstly, a method based on cross-attention is used to learn the contradictory information between images and texts from the latent space, while methods based on graph networks or contrastive learning directly model the deviations in emotion and semantics. This not only improves the detection performance but also better handles complex sarcastic situations, such as taking into account large-span semantic deviations and small-amplitude emotional reversals.

3. Incongruity Modeling Method Based on Deep Learning

Next, based on the incongruity theory, this article will review the current modeling methods for sarcasm detection using deep learning, especially large language models and multimodal large models. Different from previous works that treated sarcasm detection as a text-image classification task, in recent years, researchers have placed greater emphasis on explicit modeling of conflicts and explainable reasoning processes. In the following, this article will mainly introduce five representative directions.

3.1 Inference Agent Architecture Based on “World Model”

One important type of work no longer views sarcasm detection as a problem of pattern matching, but rather restructures it as a cognitive reasoning process. The core assumption is that sarcasm arises from the disconnection between the “literal meaning” and “social common sense or normative expectations.”

Based on the above ideas, this paper mainly focuses on multi-agent methods such as WM-SAR. These methods use a large LLM to generate different specialized agents to determine: literal meaning, contextual information, normative expectations, and speaker intent.

The gap between “literal evaluation” and “normative expectation” is used as a numerical “incongruity score” as the output. This result, along with other inference signals generated by the large model, is then fed into a small-scale logistic regressor to obtain the probability value of sarcasm. In this way, we retain the reasoning ability of the large model while improving the interpretability of its judgment process.

3.2 “Contradiction Perception” Fine-Tuning of Multimodal Large Models

Furthermore, the researchers also noted that most VL models (such as CLIP or its derivatives) are specifically trained for image-text alignment tasks, which are fundamentally different from the contradiction detection task that satire detection focuses on. Therefore, some fine-tuning paradigms based on contradiction perception were proposed.

Researchers have also pointed out that most vision-language (VL) models, such as CLIP and its variants, are originally designed for image-text matching rather than sarcasm detection. Because of this difference, several studies have explored fine-tuning strategies that explicitly consider contradiction between the two modalities.

SCARF is one example of this type of method. It introduces a clue fusion framework based on signature-constrained question-answering tasks and generates two kinds of clues, including coarse scene clues and fine-grained local clues. These clues are then provided to the large language model to help it pay more attention to the inconsistent parts of the text and the image.

The framework also combines context retrieval with local multi-view encoding. This allows the model to make use of small visual details, such as facial expressions or background text, which often play an important role in understanding sarcastic meaning.

3.3 Structured Thinking Chain and Course Learning

In response to the lack of interpretability in existing models, we aim to guide the models to simulate the human mechanism of sarcasm detection: the process of thinking and judgment is a promising direction. Structured thinking chains and curriculum learning fall under this category of work.

To train the aforementioned mechanism, InterARM proposes a three-stage training process: the first stage is the learning of the sarcasm classifier; the second stage is the learning of structured reasoning, which is used to generate reasoning chains and explicitly analyze the emotional polarity and semantic contradictions; the third stage is based on the reinforcement learning process, where the reasoning process is further optimized through selective, step-by-step, hierarchical rewards.

Furthermore, in datasets such as MSD-CoT, the reasoning process was manually annotated, forcing the model to learn reasoning links such as "the textual and graphical emotional polarities are opposite, so it is satire." This is of great help in improving the interpretability of sarcasm detection.

3.4 Generative Safety Assessment and Intent Recognition

When we expand the scope to the areas of content security and understanding of intentions, sarcasm is not merely a linguistic issue; it is often also a matter of implicit attitudes. "Using positive words in a negative context" or "telling the opposite of what is said" are very common situations. Therefore, the sarcasm judgment model should also be a generative judgment, rather than a simple binary classification result.

Specifically, the existing models can achieve the following: generate natural language explanations (for example, "This sentence is a rhetorical question and does not conform to the facts, and it belongs to derogatory sarcasm"); handle semantic inconsistencies across multiple sentences in multi-round conversations by leveraging the long context window of large models; and determine whether a language with a specific cultural background contains sarcasm based on multilingual training materials (for instance, self-mockery of the "King of Rankings" (*juanwang*) in the Chinese context).

This direction is of great significance for the content review and public opinion analysis in actual social media applications.

3.5 Unified Representation of Lightweight Visual Language Models

Additionally, during the process of launching the model, the issue of deployment speed also needs to be taken into consideration. Not all tasks require large models for inference. Therefore, some work has begun to explore lightweight VLMs for sarcasm detection. The key technical features of these methods include the following two points.

First, the unified distribution space is adopted, using sparse attention for alignment instead of direct connection; second, hierarchical reasoning is employed, comprising five stages: shallow understanding - deep semantic expression - consistency evaluation - common sense judgment - intention analysis.

This design also demonstrates excellent robustness and generalization ability in the identification of metaphorical sarcasm and implicit contrast sarcasm.

4. Key Challenges and Future Directions

This paper reviews the related work on deep sarcasm recognition from the perspective of inconsistency, and traces the development process of sarcasm recognition from shallow fusion to deep incongruous representation. It can be seen that sarcasm is often accompanied by incongruity between expectations and information, and this psychological process of incongruity is effectively formalized into a mathematical model. The deep learning method based on the incongruity at both the factual and emotional levels can effectively reflect the conflict between images and text, thereby inferring the sarcastic meaning hidden behind the text, providing a solid foundation for the study of incongruity and guiding the design of algorithms. Although these models have achieved some success on some standard datasets, they still suffer from poor interpretability, difficulty in local alignment, and lack of background information in complex contexts and semantic contradictions.

In summary, future research can further develop sarcasm recognition methods based on incongruity theory, making them more cognitive and multimodal, mainly in three aspects:

First, future studies may pay more attention to the combination of data and knowledge in multimodal sarcasm detection. Although current models have achieved good performance based on large-scale data, they still have difficulties in understanding sarcastic expressions that depend on cultural background or common sense. Therefore, introducing external knowledge sources, such as knowledge graphs (e.g., ConceptNet and VisualGenome), and incorporating common sense reasoning may help improve the understanding of these cases. With the development of large language models, another possible direction is to explore how their world knowledge can be used in sarcasm detection. Such models may not only identify the inconsistency between images and texts, but also provide a better understanding of the reasons behind this inconsistency.

Second, more efforts are needed to study fine-grained interactions between different modalities. Existing methods often calculate incongruity based on the overall similarity between an image and a sentence, which may overlook some small but important clues. For example, a certain object in an image may conflict with a specific word or description in the text. Future work could combine object detection and visual grounding techniques to explore more detailed alignment between visual and textual information. By focusing on local regions and specific attributes, models may be able to detect more subtle semantic conflicts and emotional changes.

Finally, model interpretability is still an important issue for the practical use of sarcasm detection systems. Although deep learning models can achieve high accuracy, it is often difficult to understand how they reach their decisions. Future research could further investigate explainable artificial intelligence methods in this area, especially methods that can provide human-readable explanations. For example, the model could explain which part of the image conflicts with which words in the text and why such inconsistency may indicate sarcasm. This may improve users' confidence in the system and also provide a better connection between computational models and theories of sarcasm understanding.

References

[1] Kužina, V., Petric, A. M., Barišić, M., et al. (2023). CASSED: Context-based approach for structured sensitive data detection. Expert Systems with Applications, 223, Article 119924.

[2] Fu, C. Y., Chen, Z. K., & Pan, L. (2023). A sarcasm detection method based on the fusion of textual semantics and social behavioral information. Chinese Journal of Network and Information Security, 9(4).

[3] Wen, C., Jia, G., & Yang, J. (2023). Dip: Dual incongruity perceiving network for sarcasm detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2540–2550).

[4] Li, S. X., Hu, H. J., & Liu, M. F. (2023). Social media multimodal sentiment analysis based on semantic consistency. China Sciencepaper, 18(3).

[5] Yue, T., Mao, R., Wang, H., et al. (2023). KnowleNet: Knowledge fusion network for multimodal sarcasm detection. Information Fusion, 100, Article 101921.

[6] Sun, S. Q., Wei, G. Y., & Wu, S. (2025). Sentiment analysis method integrating hypergraph enhancement and dual contrastive learning. Journal of Computer Engineering & Applications, 61(22).

[7] Vedhaviyassh, D. R., Sudhan, R., Saranya, G., et al. (2022). Comparative analysis of easyocr and tesseractocr for automatic license plate recognition using deep learning algorithm. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 966–971). IEEE.

[8] Long, Y. C., Gou, Z. N., Chen, Y. X., et al. (2025). Multi-task generative reconstruction dialogue emotion recognition based on large language models. Application Research of Computers, 42(7).

[9] Pfeifer, E., Wulf, H., Metz, K., et al. (2024). Correlations between meaning in life and nature connectedness: German-language validation of two topic-related measures and practical implications. Spiritual Care, 13(3), 242–256.

[10] Tan, Y. Y., Chow, C. O., Kanesan, J., et al. (2023). Sentiment analysis and sarcasm detection using deep multi-task learning. Wireless Personal Communications, 129(3), 2213–2237.

[11] Šandor, D., & Bagić Babac, M. (2024). Sarcasm detection in online comments using machine learning. Information Discovery and Delivery, 52(2), 213–226.

[12] Luo, J., Li, Y., Li, X., et al. (2025). PKME-MLM: A novel multimodal large model for sarcasm detection. Computers, Materials, & Continua, 83(1), 877.

Funding
This research received no external funding.

Conflicts of Interest
The authors declare no conflict of interest.

Acknowledgment
This paper is an output of the science project.