Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SYSTEMATIC REVIEW OF TEXT-IMAGE MULTIMODAL SARCASM DETECTION: MODELING APPROACHES, BENCHMARK RELIABILITY, AND EVALUATION CHALLENGES

CHAO BING[1], AZREEN AZMAN[2], NURUL AMELINA NASHARUDDIN[3] AND MASRAH AZRIFAH AZMI MURAD[4]

[1,2,3,4] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

E-mail: [1]gs73318@student.upm.edu.my, [2]azreenazman@upm.edu.my, [3]nurulamelina@upm.edu.my, [4]masrah@upm.edu.my

ABSTRACT

The increasing use of images and text to convey ideas in social media has drawn researchers' attention to the task of multimodal sarcasm detection. Multimodal sarcasm detection is different from the traditional text-only sarcasm detection in that the incongruities between the modalities have to be captured as well as the meaning of each modality. While recent years have seen some advances, there are currently no uniform standards for modelling strategies, dataset construction and evaluation. Hence this paper conducts a systematic review of the available literature in the social media context. Specifically, this research focuses on induction and classification of the current literature on fusion-based architectures, incongruity modeling strategies, knowledge guided methods and task variants other than binary classification. Moreover, we discuss popular datasets and their drawbacks like annotation noise, shortcut learning bias and lack of diversity. Meanwhile, unfair comparisons due to differences in evaluation metrics and experimental setting is discussed. The results of this study further highlight the major challenges of multimodal sarcasm detection such as subjectivity, context dependency, cross-platform variability, etc., and provide suggestions for future research on the development of more robust, interpretable, and generalizable multimodal sarcasm detection models. The overall purpose of this literature review is to structure the reference for research in this area and standardize the evaluation practice.

Keywords: *Multimodal Sarcasm Detection, Cross-Modal Interaction, Incongruity Modeling, Social Media Analytic, Systematic Literature Review*

1. INTRODUCTION

Sarcasm is a rhetorical tool that is widely used in the network communication environment [1]. It can be used to express criticism, humor or irony, typically with an indication of the discrepancy between literal meaning and real meaning [2], this is a feature that presents challenges for computational models [3].

This challenge has been magnified in the social media world. Social media is one of the most important means of human expression and public discussion in the last decade [4]. Social media content is generally spontaneous and informal, and is becoming increasingly multimodal, typically fusing text with images, emojis or memes, with the exception of traditional communication channels [5]. Sarcasm is often conveyed through the interplay between text and image, but it can also be conveyed in a single modality. When a picture of a thunderstorm is placed next to the sentence "Today is a perfect day for a picnic", the sarcasm is mostly in the contrast between the words and the picture.

While multimodal sarcasm detection can have audio, video or other modality, most of the time the study and most public datasets are related to the text-image modality, notably, that of posts and memes on social media. Thus, this work focuses on the problem of sarcasm detection based on text-image content, as it is the most studied and consequential scenario.

Although there is some advancement in recent years, multimodal sarcasm detection is still a challenge, and some problems still hinder the development of robust and generalizable models.

1.1 Subjectivity

Sarcasm is subjective and reflects the common culture background or contextual knowledge, which causes inconsistencies of labelling across datasets [6][7].

1.2 Short And Noisy Content

Social media posts are often brief and informal, using slang, emojis, and changing meme styles, and are difficult to interpret [8].

1.3 Cross-Platform Variation

Sarcasm is different across platforms, especially in text-image interactions, and this difference leads to limited models' generalizability when trained on certain datasets.

1.4 Data Limitations

Current datasets tend to be limited in size, monolingual, and mainly text-image pairs, which hinders scalability and diversity [9].

1.5 Context Dependence

Understanding of sarcasm requires additional information such as conversation context or information about the user, which cannot be modeled easily[10].

1.6 Inconsistent Evaluation Standards

Different research uses different evaluation metrics and experiments, which makes comparison very hard.

The challenges discussed above highlight the need for a systematic examination of existing approaches, datasets, and evaluation practices in multimodal sarcasm detection. Unlike previous surveys [11], [12], this review jointly examines modeling approaches, benchmark reliability, and evaluation consistency to identify gaps affecting the reliability and comparability of existing methods.

To address this gap, this study conducts a systematic literature review (SLR) of text–image multimodal sarcasm detection studies published between 2015 and 2025, following the PRISMA framework. The findings aim to provide a structured understanding of current research trends and identify challenges affecting the reliability and comparability of multimodal sarcasm detection systems.

This study focuses on tackling the following research questions:

RQ1: What methods are proposed for multimodal sarcasm detection in text-image data, and what is their approach to handling cross-modal interactions?

RQ2: What are the datasets used in multimodal sarcastic expression recognition, and what are their features and shortcomings?

RQ3: Which metrics are applied for evaluation purposes and how do inconsistencies in them impact the comparability of the results?

The main advantages of our study include the following:

1) Our research is based on a systematic review of multimodal detection of sarcasm within the framework of text-image interaction in social media.

2) We proposed the classification of different methods as fusion-based, incongruity modeling, knowledge-enhanced, and task-variant approaches.

3) We carefully analyzed existing datasets extensively utilized in practice and pointed out their limitations, such as noisy annotations and shortcut bias.

4) We detected inconsistencies in benchmarking practices and suggested solutions to that problem.

5) We summarized the key research gaps and outline future directions for achieving more robust and interpretable multimodal sarcasm detection.

2. THE SYSTEMATIC LITERATURE REVIEW PROCESS

The study conducts an SLR using the PRISMA protocol [13]. Figure 1 summarizes the review design, including five stages: database identification and search, duplicate removal, title and abstract screening, full-text eligibility assessment and quality assessment and data extraction.

2.1 The Research Protocol

This section details the search strategies employed for the comprehensive and systematic retrieval of relevant research papers. Four of the most widely used databases in the field of computer science were selected for this purpose:

- IEEE Xplore

- Scopus
- ACM Digital Library
- Google Scholar

In addition, appropriate search keywords were developed to retrieve a large number of studies that meet our inclusion criteria on multimodal sarcasm detection in social media. A preliminary investigation was conducted to identify the keywords commonly used in this research field. Accordingly, we used the main terms "sarcasm detection", "multimodal", and their respective synonyms. It should be noted that, in order to avoid missing relevant literatures, the search query "in social media" was not included. The search keywords used in this study are as follows:

("sarcasm" OR "sarcastic" OR "irony" OR "ironic" OR "satire") AND ("detection" OR "recognition" OR "identification") AND ("multimodal" OR "multi-modal" OR "cross modal" OR "cross-modal" OR "image-text")

Furthermore, as described in Section 2.2, we established specific inclusion and exclusion criteria to screen the search results. Subsequently, we defined quality assessment criteria, as shown in Section 3.2, to evaluate the quality of the studies included in this review.

2.2 Inclusion and exclusion criteria

To ensure the rigor and relevance of the studies included in this review, a set of explicit inclusion and exclusion criteria was established. The inclusion criteria aim to include studies that directly contribute to the understanding of multimodal sarcasm detection, while the exclusion criteria helped exclusive works that are outside the scope of this study. The detailed criteria are shown in Tables 1 and 2, respectively.

Table 1: Inclusion Criteria.

Criteria	Description
1	The research topic is directly related to sarcasm detection, irony detection, or satire detection.
2	The study employs multimodal information (text + image).
3	The research focuses on social media posts.
4	The study is published in a peer-reviewed journal paper, short paper, or workshop conference paper.
5	The paper is written in English.
6	Publication period: 2015-2025

Table 2: Exclusion Criteria.

Criteria	Description
1	Studies unrelated to sarcasm detection (e.g., pure humor detection, non-sarcastic sentiment analysis).
2	Single-modality studies (text-only or image-only).
3	Studies not involving natural language (e.g., pure image emotion recognition).
4	Papers with no full text available or not published.
5	Papers with no clear information of methodology.

2.3 Quality assessment criteria

Each paper was subjected to a quality assessment to ensure its high quality and relevance. Table 3 outlines the quality assessment criteria adopted in this SLR. After the practical screening phase, each full-text paper was evaluated against all the predefined quality assessment criteria listed in Table 3. A score of 1 was assigned to papers that satisfied a given criterion, while those that did not were assigned a score of 0. Only studies meeting all five criteria (with a total score of 5) were included in the final selection. Consequently, the included studies were required to focus on multimodal sarcasm detection on social media, describe the use of machine learning algorithms, integrate datasets, and present evaluation result.

Table 3: Quality assessment criteria

Criteria	Description	Evidence Indicator
Q1	Is the purpose of this paper related to multimodal sarcasm detection on social media	Address multimodal sarcasm detection in social media
Q2	Are machine learning algorithms used in the experimental method of the paper?	Employs machine learning techniques
Q3	Is the proposed method properly explained?	Explain algorithmic approach
Q4	Is public dataset included in the paper?	Dataset can be accessed
Q5	Are the details of the results and evaluation included in the research article so it can be replicated?	Provides experimental evaluation results

2.4 Conducting the study selection process

The study selection process began with searches conducted across four databases using the search

keyword strings detailed in Section 2.1. The retrieval of relevant research papers from the selected databases was carried out in December 2025, resulting in a total of 1,250 papers. After removing duplicates, the total number of papers was reduced to 1,015. Subsequently, the titles and abstracts were screened to determine whether the papers met the inclusion and exclusion criteria outlined in Section 2.2, leaving 182 papers. Thereafter, the quality assessment criteria described in Section 2.3 were applied, further narrowing the number of eligible full-text papers to 37. Consequently, a total of 37 studies were finally included in the SLR. The stage of paper selection is shown as Figure 1, and the distribution of publications by year is shown as Figure 2.

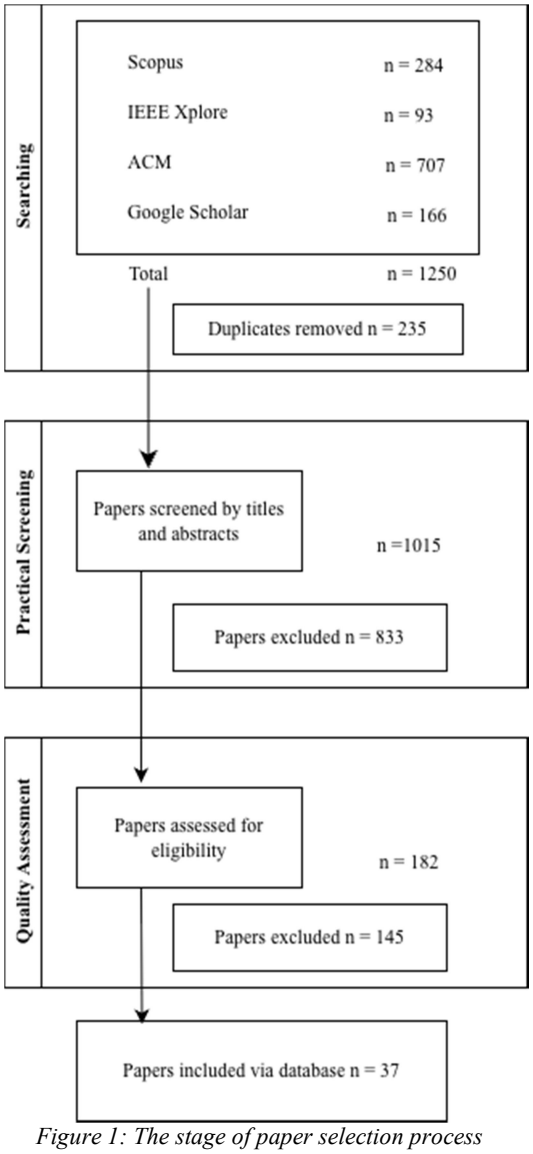


Figure 1: The stage of paper selection process

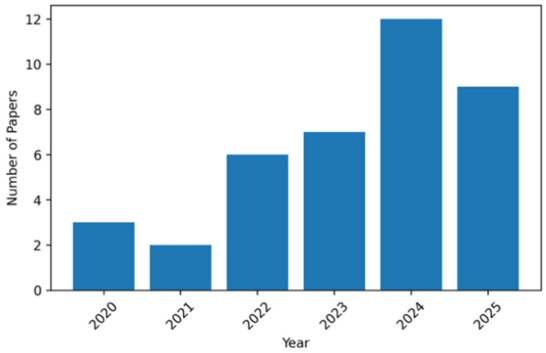


Figure 2: Distribution of publications by year

3. DATA EXTRACTION

In this section, we present a comprehensive synthesis based on an in-depth review of the 37 selected papers. In response to the research questions outlined in Section 1, we summarize the machine learning approaches proposed in these studies for addressing sarcasm detection in multimodal social media contexts. Table 4 summarizes the reviewed studies according to the proposed taxonomy, along with their representative modeling approaches.

Table 4: Categorization of reviewed studies

Category	Description	Representative Papers
Fusion-based Architectures	Encode–fuse–classify paradigm with feature-level interaction	[14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]
Incongruity Modeling Strategies	Explicit modeling of semantic or sentiment inconsistency across modalities	[29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]
Knowledge-Guided and structure-explicit modeling	Incorporation of external knowledge and structured relational modeling	[41], [42], [43], [44], [45], [46]
Task Variants beyond binary classification	Extension to fine-grained tasks such as target identification	[47], [48], [49], [50]

3.1 Fusion-based Architectures

In the early stage of multimodal sarcasm research, the mainstream method focused on shallow feature fusion: textual and visual inputs

were encoded independently and then fused by concatenation, weighted summation or light weighted attention mechanisms as a unified representation following with a simple classifier for prediction. Some early works on multimodal fusion demonstrated the advantage of utilizing complementary data from the text and images baseline compared to a single modal [14].

One of these works is the hierarchical fusion work by Cai et al. (2019), which introduced the MMSD benchmark where image properties are integrated with both textual and visual features [15].

Many other studies followed using this benchmark. Other studies used pre-trained encoder, advanced modalities, and interaction techniques to enhance the architecture of fusion model. For example, RoBERTa-ResNet uses collaborative attention mechanism and FiLM visual conditionalization [16]. BERT/CLIP-based multi-task learning and OCR-enhanced multi-stage fusion framework [17]. Multi-stage fusion approach using learning and OCR. Following this route, a number of papers have explored different approaches to fusion in order to enhance the cross-modal cross-attention and cross-routing.[18], [19], [20], [21], [22], [23], [24]. For instance, Tian et al. [25] and Ma et al. [26] Bao et al. used generative and dynamic routing approaches to achieve the balancing of weights of models [27]. Moreover, Lin and Zhu [28] used multi-granular hybrid interactive attention to enhance alignment.

Although these advances, most approaches still require a cross-modal implicit interaction, and thus, their reliability is strongly correlated with the quality of the encoders. This constraint has spurred further research aimed at directly modeling incongruity, considered as the key property of sarcasm detection.

3.2 Incongruity Modeling Strategies

Incongruity modelling is a set of techniques that involve an inconsistency between modalities, for example between the expressed emotion and the perceived emotion, or between a statement of facts and evidence of image. While implicit learning utilizes information from conventional multimodal fusion of information, incongruity modeling regards these mismatches as the central signal of sarcasm detection.

Sentiment-enhanced learning is one direction in incongruity modelling. The idea is that polarity incongruity is the most important clue to recognizing sarcasm: If the mood, meaning, or tone of the words in the text does not fit with the mood, meaning, or tone of the image or surrounding text, the incongruity is a strong indicator of a sarcastic text. Liu et al. (2022) introduced a dual-branch sentiment-sarcasm co-training framework: A shared text-visual encoder is used to feed the two parallel branches with data, one for sarcasm detection and the other for sentiment classification. This approach enables sentiment cues to act as secondary supervision for sarcasm recognition, explicitly teaching polarity differences between modalities during training [29].

In addition to sentiment-based formulations, incongruity modeling has been applied in several ways of structuring and interacting with the model. Of these, graph-based methods are especially suited to representing inconsistencies across modalities for relations. For affective incongruity, Ou and Li (2025) devised a cross-modal mapping and graph convolutional method to leverage textual polarity expressions and visual emotional cues to identify incongruities between them, which includes sarcasm instances like positive text and negative imagery (and vice versa) [30]. Lu et al. (2024) also broke down incongruity into factual conflict and emotional conflict, where they dynamically connected the two graphs with gating mechanisms to aggregate the relative importance of the two aspects at the sample level, capturing the dual causal pathways of the generation of sarcasm [31]. Li et al. (2025) extended this by proposing a multi-layered incongruity aggregation approach for perceptual ambiguity in textual, visual and modal levels. Instead of taking the incongruity as given, they employ uncertainty-weighted fusion to combat these cues and minimize the impact of weak or noisy incongruity cues [32].

Aggarwal et al. (2025) explored incongruity on a more granular level, with an alignment approach. Their method applies bidirectional cross-modal attention to identify possible conflict pairs between text and image [33]. They then imposed contrastive constraints on the candidate mismatches highlighted by these attention mechanisms, allowing the model to learn different clusters of representations for semantic consistency and semantic conflict, respectively. Compared to sentence-level or image-level coarse-grained contrast, this fine-grained alignment method at the level of word–region or phrase–object correspondences can more accurately captures the local feature of sarcasm manifestation. In addition to fine-grained alignment, dual-branch and contrastive learning paradigms have been widely adopted to explicitly enforce inconsistency

boundaries [34], [35], [36]. For example, Wen et al. [34] and Qiao et al. [40] proposed mutual-enhanced frameworks to perceive dual incongruity, while several studies leveraged sentimental constraints and global emotion perception networks to filter outlier cues [37],[38], [39].

Although incongruity modeling captures the core signal of sarcasm, it is highly sensitive to noise or ambiguous cues in social media data. Moreover, many approaches assume the existence of clear polarity conflicts, but this may not hold true in subtle or context-dependent sarcasm cases.

3.3 Knowledge-Guided and Structure-Explicit Modeling

Multimodal sarcasm detection: Knowledge and knowledge-enhanced research are intended to provide factual priors, commonsense structures, and interpretable reasoning mechanisms to better detect factual or affective inconsistencies between "what the text states" and "what the image implies."

At present, interpretable approaches based on logical reasoning that map multimodal evidence to explicit rules or logical patterns, the system not only decides whether an input is sarcastic, it also provides auditable chains of evidence, for instance, entailment or contradiction relations between text-based assertions and visual entities, thus making the question "why is it sarcastic?" read and verifiable as a reasoning process [41]. Such research is highly interpretable and domain robust, however it is restricted by the rules of the scope and rules of noise, particularly in the social media world saturated with memes and metaphors, where rule-based systems struggle with full representation. To overcome this rigidity and noise, researchers often couple with weakly supervised or semi-supervised learning with neural representation learning.

Another strand of knowledge-led research is on minimizing redundant or misleading information. Zhu et al. (2024) suggested a disentangled multi-grained distillation approach by separating multimodal features into modality-invariant and modality-specific ones. The model then shares valuable information between various levels of granularity such as local features, global features, textual units and visual regions [42]. This design will assist the model to shift attention away from superficial patterns to more stable cross-modal patterns.

Multi-granularity fusion methods have similar motivations. These methods differ from treating text and image as two global representations; rather, they model interactions on three different levels: words, phrases, objects, attributes, and scenes. This enables that the model can reflect more complex forms of consistency and contradiction between modalities [51]. These types of designs, however, also introduce model complexity, and can be prone to overfitting, particularly with limited or noisy data.

There are also graph based methods of making the cross-modal relationships more explicit. These models can represent the textually and visually presented elements as nodes, and their semantic, affective, or factual relations as edges. Liang et al. (2024) presented a multimodal graph contrastive learning model that learns both fusion and discrimination by comparing graph-level and subgraph-level representations [43]. In a similar fashion, Wei et al. (2024) used label-aware graph contrastive learning, along with back-translation augmentation. Their method first constructs unimodal graphs and then integrates them into multimodal graphs, allowing samples with the same label to be represented more closely while separating samples with different labels [44]. Similar explicit relational mapping can also be found in knowledge-enhanced hierarchical congruity models [45] and dynamic graph convolutional networks [46].

Overall, knowledge-guided and graph-based methods make the reasoning process more explicit than simple fusion models. They are useful for modeling structured relations and reducing shallow correlations, but they also introduce new challenges. They can be impacted by noisy external knowledge, incomplete graph construction, and increased computational price. Thus, an effort to make them more scalable and robust is still important research avenues

3.4 Task Variants Beyond Binary Classification

Unlike conventional binary multimodal sarcasm detection, Multimodal Sarcasm Target Identification (MSTI) requires models to not only predict whether sarcasm exists but also identify the specific textual and visual targets involved. Therefore, MSTI moves sarcasm detection from sample-level classification to fine-grained evidence localization, such as detecting sarcastic text spans and image regions.

Chen et al. (2024) proposed CofiPara, a coarse-to-fine framework for MSTI [47]. In the coarse stage, large multimodal models are used to generate rationales during sarcasm detection pretraining.

Paper	Dataset	Size(Trian/ Dev/Test)	Task Type	Language
[15], [52]	MMSD/ MMSD2.0	19,816/ 2,410/ 2,409	Binary sarcasm detection	English
[53]	Memotion	6,992/ 1,000/ 1,879	Multi-label sentiment	English
[17]	Multibully	4,097/ 585/ 1,172	Cyberbully + sarcasm	Hinglish
[54]	SarcNet	1,998/ 671/ 666	Binary sarcasm detection	English & Chinese

These rationales provide weak guidance about where potential sarcastic cues may appear. In the fine stage, the model is further trained with MSTI annotations to localize sarcasm targets at the token and region levels. This design is useful because fine-grained MSTI annotations are costly to obtain, while coarse sarcasm labels are relatively easier to use.

Besides rationale-based learning, relational modeling can also support fine-grained sarcasm understanding. For example, Wang et al. (2025) proposed RCLMuFN, which models relations such as entity–entity, entity–attribute, and word–object interactions through relational context learning and multiplex fusion [48]. Although this method is designed for multimodal sarcasm detection, its emphasis on structured cross-modal relations is relevant to target identification, because sarcasm targets often depend on specific conflicts between textual expressions and visual entities. Other notable expansions of task variants include comment-aware self-training [49] and multi-modal target localization frameworks in social tweets [50].

Overall, task variants such as MSTI extend multimodal sarcasm detection beyond binary classification. They make the model interpretable through making them find the ground of the sarcasm; however, they present new challenges regarding fine-grained annotation, more difficult evaluation metrics, and localization skills.

4. DATASETS

Benchmarking datasets have important applications in the domain of multimodal sarcasm detection. These benchmarking datasets, used in this study, differ from conventional benchmarking sarcasm datasets that have text only. This is because multimodal benchmarking datasets also contain visual content since it is part of the multimodality of sarcasm in social media. This section highlights some of the popularly used benchmarking datasets with special focus on their creation and evaluation processes, and the summary of the dataset is shown as Table 5.

Table 5: Dataset Summary.

4.1 MMSD And MMSD 2.0

The MMSD dataset [15] One of the earliest and most widely used benchmark sarcasm detection datasets. It houses English tweets, images, and auto-gleaned attribute words (including object labels, scenes and visual emotions). The dataset was initially created with weak supervision and then partially cleaned and annotated by humans. Each occurrence classified as sarcastic or non-sarcastic.

But later research has shown that MMSD has certain drawbacks, notably the presence of shortcuts like the hashtags and sarcasm-related emojis, which may cause the model to fall into superficial patterns instead of understanding both modalities. To solve this problem, MMSD 2.0 [52] A debiased and more reliable version of the original dataset was introduced.

Specifically, MMSD 2.0 eliminated tokens that are highly predictive of class labels to reduce shortcut bias and re-labeled negative examples by crowdsourcing with a three-way labeling scheme (Sarcasm, Not Sarcasm, Undecided), adjudicated by experts. The final annotation results gave a Cohen's κ of 0.811, which is considered substantial agreement. These improvements helped a lot in improving the quality of the annotations and made MMSD 2.0 more reliable as a benchmark dataset for multimodal sarcasm detection models.

Despite the numerous improvements in MMSD 2.0, the main reasons why MMSD continues to be the benchmarking dataset in such research are due to its popularity and comparability with prior works. In this context, numerous recent studies report findings for MMSD and present findings for MMSD 2.0, presenting the balancing act between consistency in the benchmark and reliability in the evaluation. This co-existence suggests a problem in the multimodal sarcasm detection space - namely that older benchmarks are not helping the field move forward, and that debiased datasets would help to facilitate reliable evaluation.

4.2 Memotion

The Memotion dataset [53] It is concerned with doing multimodal sentiment analysis of online memes having text. It contains around 10,000 English memes collected from the Internet. Sentiment polarity, sentiment type (along with sarcasm), and sentiment strength were recorded for each meme. Text component of the memes was converted using optical character recognition and further corrected using crowdsourcing. Five distinct annotators were assigned to each meme, and their majority decision was taken for final labeling.

Three benchmarks' tasks are present in the Memotion dataset: sentiment classification task, multi-class meme attribute classification (humorous, sarcastic, offensive, and motivational), and semantic scale prediction task.

4.3 Multibully

The Multibully dataset [17] is a multimodal for cyberbullying detection using memes, where both modalities are used. In particular, the text modality of this dataset includes mostly (Hinglish) Hybrid Hindi-English language from Twitter and Reddit.

In addition to the cyberbullying labeling, it provides additional labeling for sentiment, emotion and sarcasm to support multitask learning. All samples were labeled by three bilingual annotators, and the final label was based on the majority rule. The reported Fleiss' κ values range between 0.69 and 0.82, indicating moderate to good agreement among annotators.

4.4 SarcNet

SarcNet dataset [54] Contains a corpus of English and Chinese image-text pairs from social media. In contrast to previous datasets that relate the whole dataset to one type of sarcasm, the SarcNet dataset provides different types of sarcasm for text, images, and multimodal data.

The dataset was collected through the Twitter and Weibo platform, with sarcasm-specific keywords and hashtags. Two annotators labeled each sample, and a third annotator would be involved in the event of disagreement. Cohen's κ values between annotators are 0.9032 for text, 0.7129 for images and 0.9769 for multimodalities.

This approach allows for the modality-specific evaluation and emphasizes the significance of cross-modal interactions.

5. EVALUATION METRICS

5.1 Overview of evaluation metrics

Various metrics have been employed in the studies of sarcasm detection. Common metrics are:

5.1.1 Accuracy

In binary sarcasm classification, accuracy is often employed, however when the dataset is imbalanced, this can be compromised.

5.1.2 Precision / recall

Precision is the percentage of correctly classified sarcastic samples over all samples classified as sarcastic, and Recall is the percentage of actual sarcastic samples correctly classified through the model. These metrics are extremely helpful when dealing with imbalanced datasets, as they can easily lead to the wrong conclusions if based on accuracy alone.

5.1.3 F1-score

One of the most important metrics in sarcasm detection is F1-score which is a balance of precision and recall. The majority of studies report binary F1-score.

5.1.4 Object detection metrics (AP/AP50/AP75)

For tasks in MSTI, Average Precision (AP), AP50 and AP75 are the commonly used measures for image-based sarcasm target retrieval. The following metrics are based on COCO benchmark object detection evaluation protocols. AP50 and AP75 determine detection performance by the allowance of 50% and 75% overlap between the predicted bounding box and the ground-truth box, respectively. The greater the AP value, the more precise the localization of the sarcasm-related regions.

5.1.5 Exact match

This is one of the evaluation measures that is dedicated for the MSTI tasks. It is used for checking if the predicted sarcasm targets are exactly the same as the annotated targets.

5.2 Inconsistency of evaluation standards across studies

Evaluation techniques in the field of sarcasm detection are quite different from study to study, including in choice of metrics and task definition.

First, for the same task type (such as binary sarcasm classification), different studies present different sets of evaluation metrics. Others only depend on accuracy, which can lead to a higher degree of overfitting in an imbalanced dataset. Others have reported precision, recall, and F1 Score. Other studies rely on Macro-F1 as the main figure. The results from binary F1 and Macro-F1 cannot be directly compared because they are calculated differently.

Secondly, because of different task definitions, different evaluation protocols are used across datasets. For instance, the MMSD family of datasets rely on accuracy and F1-score for binary classification. For multi-label classification tasks, Memotion applies Macro-F1 as its evaluation metric, whereas MSTI tasks use object detection measures like AP and Exact Match. Consequently, it is impossible to compare performance between datasets.

The complex assessment framework makes establishing authentic methodological advances challenging. A better result in one of the studies can be due to the choice of metric or the properties of the dataset and not to any improvement in the model.

5.3 Implication and Recommendations for Standardized Evaluation

Lack of standardized evaluation protocols are a big challenge for the development of sarcasm detection research. Inconsistent metrics makes it difficult to compare results between studies and makes it hard to realize any true improvement in model performance. To solve these problems, we suggest:

1) F1-score and Macro-F1 should be used as a key metric for sarcasm detection, particularly when dealing with imbalanced datasets.
2) Reporting complete precision–recall–F1 statistics, not accuracy.
3) Lending clarity to the evaluation process, such as class distribution and confusion matrices.
4) Improving the reproducibility of commonly used datasets by standardizing reporting.
5) If using multimodal data, reporting results on plain text, plain image and multimodal data separately.
6) For MSTI tasks, the use of a unified evaluation protocols based on EM and AP approach.

These recommendations can improve the reliability and comparability of multimodal sarcasm detection systems, supporting both future research and practical applications in social media analysis.

6. DISCUSSION AND CONCLUSION

This work conducts SLR of Multimodal Sarcasm Detection in social media with a focus on approaches that were used in previous studies for modeling text-image interaction, creating benchmark datasets and evaluating model performance.

First, we have reviewed and analyzed various methodologies that start from early fusion-based frameworks up to complex ones such as graph-based frameworks, contrastive learning, and transformer-based structures. It is clear that the field has evolved from the use of fusion up to the inclusion of multimodal interactions and semantics inconsistencies. However, these methods still face challenges in dealing with ambiguous or context-dependent expressions.

Second, we discussed the datasets that are commonly used, and we indicated the modalities and constraints of the datasets. Although several datasets are publicly available, most are small and lack linguistic diversity. These limitations demonstrate that improving datasets is as essential as improving model architecture.

Third, we noticed that there were significant discrepancies regarding the metrics used in evaluations from study to study. These differences make it hard to compare models, as well as hide methodologies. For this reason, it is important to have a standard way of evaluating them. This highlights the importance of establishing standardized and transparent assessment methods.

Based on the findings, some of the problems related to multimodal sarcasm detection include the subjective nature of sarcasm, cultural dependency, poor diversification of datasets, and the different evaluation methods applied in various research studies.

Nevertheless, this review has several limitations. First, the literature search was restricted to four

databases and English-language publications from 2015 to 2025, which may have excluded relevant studies. Second, the review focuses on text–image sarcasm detection and does not cover other multimodal settings such as audio or video. Third, differences in datasets, evaluation metrics, and experimental configurations across the reviewed studies limited the possibility of conducting direct quantitative comparisons of model performance.

Future work should focus on constructing bigger and more varied multimodal datasets, including additional contextual data, and creating models with enhanced reasoning and interpretability capabilities. In addition, it would be important to devise common evaluation protocols and benchmarks for advancing the field in a systematic manner.

To conclude, this literature review presents a holistic understanding of the current research trends in multimodal sarcasm detection and provides some directions for future research.

REFERENCES:

[1] S. Chauhan, S. R. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2020, pp. 4351–4360. doi: 10.18653/v1/2020.acl-main.401.

[2] D. Ding et al., "Multi-Modal Sarcasm Detection with Prompt-Tuning," 2022, doi: 10.1109/ACAIT56212.2022.10137937.

[3] O. Vitman, Y. Kostiuk, G. Sidorov, and A. Gelbukh, "Sarcasm detection framework using context, emotion and sentiment features," Expert Syst. Appl., vol. 234, p. 121068, Dec. 2023, doi: 10.1016/j.eswa.2023.121068.

[4] R. M. Bond et al., "A 61-million-person experiment in social influence and political mobilization," Nature, vol. 489, no. 7415, pp. 295–298, Sep. 2012, doi: 10.1038/nature11421.

[5] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of Emojis," PLoS One, vol. 10, no. 12, p. e0144296, Dec. 2015, doi: 10.1371/journal.pone.0144296.

[6] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic Sarcasm Detection," ACM Comput. Surv., vol. 50, no. 5, pp. 1–22, Sep. 2018, doi: 10.1145/3124420.

[7] B. C. Wallace, D. K. Choe, L. Kertz, and E. Charniak, "Humans Require Context to Infer Ironic Intent (so Computers Probably do, too)," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 512–516. doi: 10.3115/v1/P14-2084.

[8] D. Bamman and N. Smith, "Contextualized Sarcasm Detection on Twitter," Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, no. 1, pp. 574–577, Aug. 2021, doi: 10.1609/icwsm.v9i1.14655.

[9] M. Lu et al., "A multi-modal sarcasm detection model based on cue learning," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-94266-w.

[10] N. A. Helal, A. Hassan, N. L. Badr, and Y. M. Afify, "A contextual-based approach for sarcasm detection," Sci. Rep., vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-65217-8.

[11] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic Sarcasm Detection," ACM Comput. Surv., vol. 50, no. 5, pp. 1–22, Sep. 2018, doi: 10.1145/3124420.

[12] S. Farabi, T. Ranasinghe, D. Kanojia, Y. Kong, and M. Zampieri, "A Survey of Multimodal Sarcasm Detection," 2024.

[13] R. Sarkis-Onofre, F. Catalá-López, E. Aromataris, and C. Lockwood, "How to properly use the PRISMA Statement," Syst. Rev., vol. 10, no. 1, p. 117, Dec. 2021, doi: 10.1186/s13643-021-01671-z.

[14] R. Schifanella, P. de Juan, J. R. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," Association for Computing Machinery, Inc acmhelp@acm.org, 2016, pp. 1136–1145. doi: 10.1145/2964284.2964321.

[15] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," Association for Computational Linguistics (ACL), 2020, pp. 2506–2515. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077514124&partnerID=40&md5=6d315db26566380822583f88474701d6

[16] A. B. Bhat and A. S. Chauhan, "A Deep Learning based approach for MultiModal Sarcasm Detection," Institute of Electrical and Electronics Engineers Inc., 2022, pp. 2523–2528. doi: 10.1109/ICAC3N56670.2022.10074506.

[17] K. Maity, P. Jha, S. Saha, and P. Bhattacharyya, "A Multitask Framework for Sentiment, Emotion and Sarcasm aware Cyberbullying Detection from Multi-modal

Code-Mixed Memes," in Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, in SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1739–1749. doi: 10.1145/3477495.3531925.

[18] H. Fang, D. Liang, and W. Xiang, "Multi-modal sarcasm detection based on multi-channel enhanced fusion model," Neurocomputing, vol. 578, p. 127440, 2024.

[19] X. He, L. Yu, S. Tian, Q. Yang, J. Long, and B. Wang, "VIEMF: Multimodal metaphor detection via visual information enhancement with multimodal fusion," Inf. Process. Manag., vol. 61, no. 3, 2024, doi: 10.1016/j.ipm.2024.103652.

[20] Y. Liu, Y. Zhang, Q. Li, B. Wang, and D. Song, "What Does Your Smile Mean? Jointly Detecting Multi-Modal Sarcasm and Sentiment Using Quantum Probability," Association for Computational Linguistics (ACL), 2021, pp. 871–880. doi: 10.18653/v1/2021.findings-emnlp.74.

[21] Q. Wu, W. Fang, W. Zhong, F. Li, Y. Xue, and B. Chen, "Dual-level adaptive incongruity-enhanced model for multimodal sarcasm detection," Neurocomputing, vol. 612, p. 128689, 2025.

[22] F. Yao, X. Sun, H. Yu, W. Zhang, W. Liang, and K. Fu, "Mimicking the Brain's Cognition of Sarcasm From Multidisciplines for Twitter Sarcasm Detection," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 1, pp. 228–242, 2023, doi: 10.1109/TNNLS.2021.3093416.

[23] K. N. Phan, G. Lee, H. Yang, and S. Kim, "Little Flower at Memotion 2.0 2022: Ensemble of Multi-Modal Model using Attention Mechanism in MEMOTION Analysis," in CEUR Workshop Proceedings, CEUR-WS, 2022. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128351016&partnerID=40&md5=fdaa61cdc1d69561ac475b656f26cd1c

[24] L. Song, Z. Zhao, Y. Ma, Y. Liu, and J. Li, "Global-Aware Attention Network for Multi-modal Sarcasm Detection," in Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 2409–2414. doi: 10.1109/SMC53992.2023.10394483.

[25] Y. Tian, N. Xu, R. Zhang, and W. Mao, "Dynamic Routing Transformer Network for Multimodal Sarcasm Detection," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2023, pp. 2468–2480. doi: 10.18653/v1/2023.acl-long.139.

[26] H. Ma, D. He, X. Wang, D. Jin, M. Ge, and L. Wang, "Multi-Modal Sarcasm Detection Based on Dual Generative Processes," in IJCAI International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2024, pp. 2279–2287. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204297932&partnerID=40&md5=2495131fe0900ae2ea7ce1bbfb6bb6ed

[27] Y. Bao, X. Zhao, P. Zhang, Y. Qi, and H. Li, "HIAN: A hybrid interactive attention network for multimodal sarcasm detection," Pattern Recognit., vol. 164, 2025, doi: 10.1016/j.patcog.2025.111535.

[28] J. Lin and X. Zhu, "Multi-granular dynamic interaction network for multimodal sarcasm detection," Multimed. Syst., vol. 31, no. 3, Jun. 2025, doi: 10.1007/s00530-025-01786-6.

[29] H. Liu, R. Wei, G. Tu, J. Lin, C. Liu, and D. Jiang, "Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection," Information Fusion, vol. 108, p. 102353, 2024.

[30] L. Ou and Z. Li, "Multi-modal Sarcasm Detection on Social Media via Multi-Granularity Information Fusion," ACM Trans. Multimedia Comput. Commun. Appl., vol. 21, no. 3, Mar. 2025, doi: 10.1145/3715139.

[31] Q. Lu, Y. Long, X. Sun, J. Feng, and H. Zhang, "Fact-sentiment incongruity combination network for multimodal sarcasm detection," Information Fusion, vol. 104, 2024, doi: 10.1016/j.inffus.2023.102203.

[32] K. Li, Y. Chen, Q. Wu, W. Mai, F. Li, and Y. Xue, "Ambiguity-aware Multi-level Incongruity Fusion Network for Multi-Modal Sarcasm Detection," in Proceedings - International Conference on Computational Linguistics, COLING, Association for Computational Linguistics (ACL), 2025, pp. 380–391. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85218500021&partnerID=40&md5=e78c891f32fff6834801a5a601ee8331

[33] S. Aggarwal, A. Pandey, and D. K. Kumar Vishwakarma, "Extracting cross-modal

semantic incongruity with attention for multimodal sarcasm detection," Applied Intelligence, vol. 55, no. 12, 2025, doi: 10.1007/s10489-025-06717-6.

[34] C. Wen, G. Jia, and J. Yang, "DIP: Dual Incongruity Perceiving Network for Sarcasm Detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2023, pp. 2540–2550. doi: 10.1109/CVPR52729.2023.00250.

[35] L. Ou and Z. Li, "Modeling inter-modal incongruous sentiment expressions for multi-modal sarcasm detection," Neurocomputing, vol. 616, p. 128874, 2025.

[36] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2020, pp. 3777–3786. doi: 10.18653/v1/2020.acl-main.349.

[37] B. Yu and X. Ji, "Multi-modal sarcasm detection based on emotion perception and cross-modality attention fusion," Journal of Intelligent and Fuzzy Systems, vol. 46, no. 4, pp. 8361–8374, 2024, doi: 10.3233/JIFS-233163.

[38] S. Yuan, Y. Wei, H. Zhou, Q. Xu, M. Chen, and X. He, "Enhancing Semantic Awareness by Sentimental Constraint with Automatic Outlier Masking for Multimodal Sarcasm Detection," IEEE Trans. Multimedia, 2025, doi: 10.1109/TMM.2025.3543074.

[39] Z. Xi, B. Yu, and H. Wang, "Multimodal sarcasm detection based on sentiment-clue inconsistency global detection fusion network," Expert Syst. Appl., vol. 275, 2025, doi: 10.1016/j.eswa.2025.127020.

[40] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie, "Mutual-Enhanced Incongruity Learning Network for Multi-Modal Sarcasm Detection," AAAI Press, 2023, pp. 9507–9515. doi: 10.1609/aaai.v37i8.26138.

[41] H. Liu, W. Wang, and H. Li, "Interpretable Multimodal Misinformation Detection with Logic Reasoning." [Online]. Available: https://github.com/less-and-less-bugs/

[42] Z. Zhu, X. Cheng, G. Hu, Y. Li, Z. Huang, and Y. Zou, "Towards Multi-modal Sarcasm Detection via Disentangled Multi-grained Multi-modal Distilling," European Language Resources Association (ELRA), 2024, pp. 16581–16591. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195923308&partnerID=40&md5=3f0c940324d7d86fe39bcb3d7ecad2f0

[43] B. Liang, L. Gui, Y. He, E. Cambria, and R. Xu, "Fusion and Discrimination: A Multimodal Graph Contrastive Learning Framework for Multimodal Sarcasm Detection," IEEE Trans. Affect. Comput., vol. 15, no. 4, pp. 1874–1888, 2024, doi: 10.1109/TAFFC.2024.3380375.

[44] Y. Wei, M. Duan, H. Zhou, Z. Jia, Z. Gao, and L. Wang, "Towards multimodal sarcasm detection via label-aware graph contrastive learning with back-translation augmentation," Knowl. Based. Syst., vol. 300, 2024, doi: 10.1016/j.knosys.2024.112109.

[45] H. Liu, W. Wang, and H. Li, "Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement," Association for Computational Linguistics (ACL), 2022, pp. 4995–5006. doi: 10.18653/v1/2022.emnlp-main.333.

[46] J. Hao, J. Zhao, and Z. Wang, "Multi-Modal Sarcasm Detection via Graph Convolutional Network and Dynamic Network," in Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, in CIKM '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 789–798. doi: 10.1145/3627673.3679703.

[47] Z. Chen, H. Lin, Z. Luo, M. Cheng, J. Ma, and G. Chen, "CofiPara: A Coarse-to-fine Paradigm for Multimodal Sarcasm Target Identification with Large Multimodal Models," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2024, pp. 9663–9687. doi: 10.18653/v1/2024.acl-long.522.

[48] T. Wang et al., "RCLMuFN: Relational context learning and multiplex fusion network for multimodal sarcasm detection," Knowl. Based. Syst., vol. 319, Jun. 2025, doi: 10.1016/j.knosys.2025.113614.

[49] C. Xu and J. Li, "Borrowing Human Senses: Comment-Aware Self-Training for Social Media Multimodal Classification," Association for Computational Linguistics (ACL), 2022, pp. 5644–5656. doi: 10.18653/v1/2022.emnlp-main.381.

[50] J. Wang, L. Sun, Y. Liu, M. Shao, and Z. Zheng, "Multimodal sarcasm target identification in tweets," in Proceedings of the 60th Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8164–8175.

[51] L. Ou and Z. Li, "Modeling Multi-Task Joint Training of Aggregate Networks for Multi-Modal Sarcasm Detection," in Proceedings of the 2024 International Conference on Multimedia Retrieval, in ICMR '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 833–841. doi: 10.1145/3652583.3658015.

[52] L. Qin et al., "MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2023, pp10834–10845. doi: 10.18653/v1/2023.findings-acl.689.

[53] C. Sharma et al., "SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, Stroudsburg, PA, USA: International Committee for Computational Linguistics, 2020, pp. 759–773. doi: 10.18653/v1/2020.semeval-1.99.

[54] T. Yue, X. Shi, R. Mao, Z. Hu, and E. Cambria, "SarcNet: a multilingual multimodal sarcasm detection dataset," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 14325–14335.