When Emojis Speak Louder Than Words: Multimodal Sarcasm Detection

Natural Language Processing

Vasiliki Pantelopoulou
vpantelb@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

Panagiota Nalmpanti
pnalmpaa@csd.auth.gr
Aristotle University of Thessaloniki
Thessaloniki, Greece

Abstract

Sarcasm detection is challenging because the intended meaning of a sentence often differs from its literal interpretation. In this work, we build a strong text-only baseline inspired by Khan et al., using RoBERTa as the encoder enhanced with cascaded multi-head attention and depthwise convolution blocks. We then propose a multimodal variant that extends the baseline by incorporating emoji-based visual information. Specifically, emojis are rendered into a small image grid and processed with a CLIP vision encoder, while an additional emoji-count feature is included in the fusion layer. Both models are evaluated on the iSarcasmEval dataset using a controlled train/validation/test setup. Results show that the multimodal model achieves comparable performance to the baseline, suggesting that emoji-based signals can contribute in some cases but do not consistently improve sarcasm detection.

1 Introduction

Sarcasm detection is a challenging task in natural language processing, as sarcastic expressions often convey meanings that differ from the literal interpretation of the text. In many cases, understanding sarcasm requires contextual cues, shared knowledge, or non-linguistic signals that are not explicitly encoded in words alone. As a result, even strong text-based models can struggle to correctly identify sarcastic content. With the widespread use of social media, emojis have become an important part of online communication. Emojis frequently act as emotional or pragmatic markers, reinforcing or contradicting the textual message. In sarcastic tweets, emojis are often used to signal irony, exaggeration, or mockery. However, most sarcasm detection approaches rely solely on textual information and either ignore emojis or treat them as ordinary tokens, potentially losing valuable semantic signals.

Recent advances in transformer-based language models, such as RoBERTa, have significantly improved performance in text-based sarcasm detection. Nevertheless, these models remain limited when sarcasm is expressed through a combination of text and visual or symbolic cues. This motivates the exploration of multimodal approaches that incorporate information beyond plain text.

In this work, we investigate whether incorporating emoji-based visual information can improve sarcasm detection. Following the methodology proposed by Khan et al. [1], we first establish a strong text-only baseline based on a RoBERTa architecture enhanced with attention and depth-wise convolutional blocks. Building upon this baseline, we propose a multimodal variant that combines textual representations with emoji-derived visual features extracted using a pretrained vision model. Both models are evaluated on the iSarcasmEval dataset under a controlled experimental setup to ensure a fair and meaningful comparison.

Our results show that while the text-only baseline remains highly effective, the multimodal approach can successfully capture certain sarcasm cases that depend on emoji usage. Through quantitative evaluation and qualitative analysis of prediction differences, we highlight both the strengths and limitations of emoji-aware multimodal sarcasm detection.

2 Related Work

Sarcasm detection has been extensively studied due to its dependence on implicit meaning and contextual cues. Recent approaches predominantly rely on transformer-based models such as BERT and RoBERTa, which achieve strong performance by leveraging large-scale pretraining.

Khan et al. [1], introduced a transformer-based architecture augmented with attention and convolutional blocks, showing improved sarcasm detection performance, particularly in low-resource settings. This architecture forms the basis of our text-only baseline.

Additionally, prior work has explored multimodal sarcasm detection, highlighting the role of emojis and visual cues in social media communication. However, most studies focus on explicit image–text pairs, while the contribution of emoji-derived visual representations remains relatively underexplored.

3 Dataset and Preprocessing

We conduct our experiments on the iSarcasmEval dataset, which consists of English tweets annotated for sarcasm. The dataset is split into training, validation, and test sets. The validation split is created from the training data using stratified sampling to preserve the original class distribution, while the test set is kept completely unseen during training and model selection to avoid data leakage.

The class distribution of the training set is shown in Figure 1. As observed, the dataset is imbalanced, with non-sarcastic tweets being significantly more frequent than sarcastic ones. For this reason, we report F1-score for the sarcastic class and employ a weighted cross-entropy loss during training to mitigate the effect of class imbalance.

Data preprocessing follows an adapted version of Algorithm 1 from Khan et al. [1]. Each tweet undergoes minimal text cleaning, including removal of URLs, normalization of whitespace, and replacement of common slang expressions with their standard forms. Unlike traditional NLP pipelines, no aggressive token filtering or

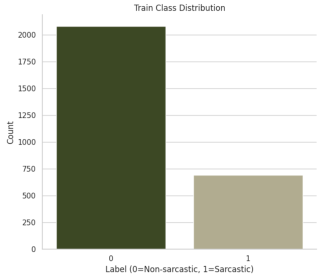

Figure 1: Class distribution of the training set in the iSarcasmEval dataset

lemmatization is applied, as the models rely on pretrained transformer tokenization.

To reduce overfitting and improve robustness, data augmentation is applied only to the training set using synonym replacement with a fixed probability. Validation and test sets are left unaugmented to ensure fair evaluation.

For multimodal modeling, the original unprocessed tweet text is preserved in parallel with the cleaned text. Emojis are extracted from the raw text and rendered as images, enabling the integration of visual emoji information in the multimodal variant. This design ensures that both text-only and multimodal models are trained and evaluated under identical preprocessing and data splits, enabling a controlled and fair comparison.

4 Methodology

In this work, sarcasm detection is treated as a binary classification task. Our methodology follows and extends the approach proposed by Khan et al. [1], who introduce a transformer-based architecture enhanced with attention and convolutional components for sarcasm detection.

We implement two models: a text-only baseline that closely follows the architecture of Khan et al. [1], and a multimodal variant that extends this baseline by incorporating emoji-based visual information.

4.1 Text-only Baseline

The baseline model is directly inspired by the architecture proposed by Khan et al. [1]. A pretrained RoBERTa encoder is used to generate contextualized token representations. On top of the encoder, we apply cascaded blocks that combine multi-head self-attention and depth-wise convolution, as described in the original paper, in order to capture both global contextual dependencies and local linguistic patterns. The final representation corresponding to the [CLS] token is used as a sentence-level embedding and passed through a linear classifier to predict sarcasm.

4.2 Multimodal Variant

Following the same text backbone proposed by Khan et al.[1], we extend the baseline model with an additional emoji-based modality. Emojis appearing in each tweet are rendered into a fixed-size image arranged in a 2×2 grid. Tweets without emojis are assigned a placeholder image to ensure a consistent visual input.

Visual features are extracted using a pretrained CLIP vision encoder. In addition, the number of emojis in each tweet is included as a scalar feature. The final representation is formed by concatenating the text features obtained from the paper-based text encoder, the visual features from the CLIP model, and the emoji count. This fused representation is then passed to a feed-forward classifier for final prediction.

To ensure a fair and controlled comparison, the text encoder of the multimodal model is initialized with the trained weights of the baseline model following Khan et al.[1], while the majority of the vision encoder remains frozen during training.

5 Experimental Setup

All experiments are conducted on the **iSarcasmEval (English)** dataset following a strict train/validation/test protocol to prevent data leakage. The original training set is split into **80% training and 20% validation** using stratified sampling, while the official test set is used only once for final evaluation.

5.1 Models

Two models are evaluated in this study. The **baseline model** is a text-only architecture based on RoBERTa, enhanced with cascaded **Multi-Head Attention and Depthwise Convolution blocks**, following the design proposed by Khan et al. [1].

The **multimodal variant** extends the baseline by incorporating emoji-based visual information. Emoji sequences are rendered into images and processed using a pretrained **CLIP Vision Transformer**. In addition, the number of emojis in each tweet is provided as a scalar feature. The final prediction is produced through late fusion of textual, visual, and scalar features.

To ensure a fair comparison, the text encoder of the multimodal model is initialized with the trained weights of the baseline model, while most of the vision encoder remains frozen during training. Only the last CLIP encoder layer is unfrozen to allow limited adaptation.

5.2 Training Configuration

Both models are trained using the **AdamW** optimizer with a learning rate of 2×10^{-5} and weight decay of 0.01. Training is performed for **4 epochs** with linear learning rate scheduling and gradient clipping.

To address class imbalance in the training data, a **weighted cross-entropy loss** is employed, with class weights computed from the label distribution of the training set.

5.3 Evaluation

Model selection is based on the **validation F1-score** of the sarcastic class. Final performance is reported on the held-out test set using **accuracy, precision, recall, and F1-score**, with particular emphasis on the sarcastic class F1-score, in line with prior work and dataset characteristics. Confusion matrices and qualitative examples are also used for analysis.

6 Results

This section presents the quantitative and qualitative results of the proposed models on the iSarcasmEval test set. We compare the text-only baseline with the proposed multimodal variant under the same experimental conditions.

6.1 Quantitative Results

Table 1 reports the performance of the baseline (text-only) model and the multimodal variant (text + emoji-image) on the iSarcasmEval test set. We evaluate both models using Accuracy, Precision, Recall, and F1-score for the sarcastic class (label=1). Since sarcasm detection is an imbalanced classification task, we focus mainly on the F1-score as the most informative metric.

Table 1: Performance comparison on the iSarcasmEval test set

Model	Accuracy	Precision	Recall	F1-score
Baseline (Text-only)	0.7850	0.3366	0.5200	0.4086
Multimodal Variant	0.8093	0.3696	0.4750	0.4158

Overall, the multimodal variant slightly improves the F1-score compared to the baseline (0.4158 vs 0.4086), showing that emoji-based visual information can provide a small benefit in sarcasm detection. The variant also achieves higher precision (from 0.3366 to 0.3696), suggesting fewer false positive sarcasm predictions. However, recall decreases (from 0.5200 to 0.4750), meaning that the multimodal model misses more sarcastic instances than the baseline. This indicates that the multimodal approach improves performance mainly by increasing precision, while the baseline remains stronger in capturing a larger portion of sarcastic tweets.

6.2 Confusion Matrix Analysis

To better understand the behavior of our models beyond aggregate metrics, we analyze the confusion matrices on the full iSarcasmEval test set for both the baseline model inspired by Khan et al. [1] and our multimodal variant. A confusion matrix highlights how predictions are distributed into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing insight into the types of errors each model makes.

Baseline (Text-only). As shown in Figure 2, the baseline confusion matrix shows that the model correctly classifies most non-sarcastic tweets, achieving **995 true negatives**. However, it predicts sarcasm in **205 non-sarcastic cases** (false positives), indicating that it sometimes overestimates sarcasm when relying only on textual cues. For the sarcastic class, the baseline achieves **104 true positives** and **96 false negatives**, which suggests that it captures a reasonable portion of sarcasm, but still misses a significant number of sarcastic examples.

Multimodal Variant (Text + Emoji Visual Features). As shown in Figure 3, the multimodal variant improves classification of non-sarcastic tweets by increasing true negatives to **1038** and reducing false positives to **162**. This indicates that the multimodal model becomes more *conservative* when predicting sarcasm, meaning that it predicts sarcasm less frequently unless it is more confident. For the sarcastic class, the multimodal model achieves **95 true positives** and **105 false negatives**, which shows a slightly lower ability to detect sarcasm compared to the baseline.

Interpretation. Overall, these results confirm a clear **precision–recall trade-off**. The baseline model detects slightly more sarcastic tweets (higher recall), but at the cost of more false positives. In contrast, the multimodal model reduces false positives (higher precision), but misses more sarcastic tweets (lower recall). This suggests that emoji-derived visual features can influence the final decision boundary, making the model stricter in predicting sarcasm, but the benefit is not consistent across all sarcastic cases.

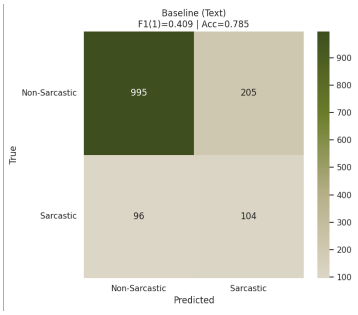

Figure 2: Confusion matrix of the baseline text-only model on the iSarcasmEval test set.

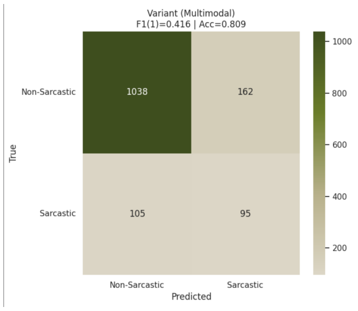

Figure 3: Confusion matrix of the proposed multimodal variant (text + emoji-based visual features) on the iSarcasmEval test set.

6.3 Qualitative Analysis

To further analyze model behavior beyond quantitative scores, we examine representative test examples where the baseline model and the multimodal variant disagree.

Golden Examples. Golden examples correspond to cases where the baseline predicts *non-sarcastic* (0), but the multimodal variant correctly predicts *sarcastic* (1). These cases suggest that the additional emoji-related modality can provide complementary cues that shift the model towards the correct interpretation.

Examples include:

- *"I just can't wait to spend time with my family over Christmas! I just love being the only single one and the many many questions asking when will I get a boyfriend face-with-rolling-eyes"*
  (**True=1**, Baseline=0, Variant=1)

- *"Top 10 pools in my book: 1. Swimming pool 2. Paddling pool 3. Above-ground pool 4. Family pool 5. Architectural pool 6. Indoor pool 7. Lap pool 8. Olympic size pool 9. Natural pool 10. Salt water pool Sorry Liverpool you are not top 10 pools in my book loudly-crying-faceloudly-crying-faceloudly-crying-face"* (**True=1**, Baseline=0, Variant=1)

Even when explicit emojis are not always present in these particular samples, the multimodal branch still affects the final decision boundary and helps correct some difficult sarcastic cases.

Reverse Golden Examples. Reverse golden examples are cases where the baseline correctly identifies sarcasm, but the multimodal variant predicts *non-sarcastic*. This highlights that emoji-based information does not always improve detection and may introduce noise or distract the model when sarcasm is mainly expressed through text.

Examples include:

- *"yeah man just walk right in. clearly you can read the sign that says "only two patients inside at once" when there are already four people in here cant you"* (**True=1**, Baseline=1, Variant=0)
- *" I'm heading out to snag my COVID-19 inoculation! Thanks government mandate! Super pleased to be doing my part even though the vaccines do not stop one from getting the disease, they do not stop the spread, and it leads to a life of non-stop useless boosters! Woo-hoo!"* (**True=1**, Baseline=1, Variant=0)

Overall, this qualitative analysis confirms that the multimodal variant can improve performance on specific samples, but its contribution is not consistent across all sarcastic tweets. This aligns with the quantitative results, where the multimodal model achieved a small improvement in F1-score but did not uniformly outperform the baseline across all cases.

7 Coclusion

The goal of this project was to examine whether adding emoji-based visual information can improve sarcasm detection, compared to a strong text-only baseline inspired by Khan et al. [1]. For this reason, we implemented two comparable models under the same experimental setup: a text-only baseline and a multimodal variant that additionally uses an emoji-image representation extracted with CLIP.

Overall, the multimodal variant achieved slightly higher performance than the baseline in terms of F1-score for the sarcastic class. Specifically, the baseline obtained an F1-score of 0.409, while the multimodal model reached 0.416. This indicates that emoji-related information can provide a small benefit, but the improvement is limited.

A more detailed look at the results shows an important precision–recall trade-off. The multimodal model improved precision (0.370 vs. 0.337), meaning that when it predicts sarcasm it is more often correct. However, the baseline achieved higher recall (0.520 vs. 0.475), meaning that it detects more sarcastic tweets overall. This behavior is also confirmed by the confusion matrices: the multimodal model produces fewer false positives but slightly more false negatives compared to the baseline.

The qualitative analysis further supports this observation. In multiple golden examples, the multimodal model correctly detected sarcasm when the baseline failed. At the same time, reverse golden cases show that the multimodal branch is not always beneficial, especially when sarcasm is mainly expressed through linguistic irony rather than emoji signals. This suggests that emojis can contribute useful cues in certain contexts, but they are not a reliable indicator of sarcasm across all samples in the dataset.

Finally, the limited improvement can be explained by the characteristics of iSarcasmEval. Only a subset of tweets includes emojis, meaning that multimodal features are not available for all examples. Additionally, sarcasm is often expressed through complex text patterns, context, and pragmatic meaning, which may not be captured strongly by emoji visual representations alone.

In summary, our results suggest that emoji-based multimodal sarcasm detection is a valid and meaningful extension of the baseline, but it does not guarantee large gains. The multimodal approach mainly shifts the model towards more conservative sarcasm predictions, improving precision but slightly reducing recall.

8 Limitations

This work has several limitations. First, emojis are not present in all tweets of the iSarcasmEval dataset, so the multimodal branch cannot contribute equally to every sample. Second, emojis were represented only through a simple rendered image grid, which may lose semantic information such as emoji meaning, sequence, or interaction with the text. Third, the CLIP vision encoder was mostly frozen during training, which may limit the ability of the model to adapt visual emoji features to sarcasm detection. Finally, sarcasm often depends on deeper context and world knowledge, which is not available in this dataset and cannot be fully captured by text and emojis alone.

9 Future Work

Future work could explore stronger emoji modeling approaches, such as using dedicated emoji embeddings or learning emoji semantics jointly with text. Another improvement would be to fine-tune more layers of the CLIP encoder or use a lighter vision model trained specifically on emoji data. In addition, applying more advanced fusion mechanisms (e.g., cross-attention between text and emoji features) could improve interaction between modalities. Finally, testing the approach on other sarcasm datasets or multilingual settings would provide a more robust evaluation of the multimodal contribution.

References

[1] Shumaila Khan, Iqbal Qasim, Wahab Khan, Khursheed Aurangzeb, Javed Ali Khan, and Muhammad Shahid Anwar. 2025. A novel transformer attention-based approach for sarcasm detection. *Expert Systems* 42, 1 (2025), e13686.