Multimodal Irony/Sarcasm Detection

Beyza Nur Koç
koc.beyza@metu.edu.tr
Middle East Technical University
Ankara, Turkey

ABSTRACT

This part is initially left blank.

KEYWORDS

Multimodal Irony Detection, Sarcasm Detection, Large Language Models, Multimodal LLMs, Qwen-VL, RoBERTa, Social Media, Turkish NLP

ACM Reference Format:
Beyza Nur Koç. 2025. Multimodal Irony/Sarcasm Detection. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Irony is the difference between what someone says and what they really mean. It is often used to criticize or make fun of a situation, especially when something bad or unwanted happens. In daily life, many people share their thoughts and feelings on social media. Sometimes they use irony or sarcasm to make a difficult situation feel easier or more humorous. Because of this, detecting irony has become an important research topic in Natural Language Processing (NLP), especially with the help of Large Language Models (LLMs).

In some studies, researchers have used different models to detect irony and sarcasm. Some of them focused only on text [1], while others used only images [2]. However, these single-modality approaches are often not enough to understand the full meaning. A person may not show irony clearly through only text or only image, but when both are combined, the contrast becomes more visible. That is why multimodal approaches, which use both text and image together, are becoming more popular. Some research suggests that combining modalities leads to better performance compared to using text or image alone [2].

Studies that use multimodal approaches for irony detection generally follow two main strategies for multimodal irony detection. Some use multimodal language models that process both text and image together as a single input. For example, the work by [4] employs a unified transformer-based model that receives both modalities at once and learns joint representations. On the other hand, other studies use separate models for each modality and combine their outputs later. For instance, [3] uses a CNN to extract visual features and an LSTM for textual features, which are then merged using a fusion layer. Both approaches are more effective than using only one modality. Unified vision-language models are trained to understand the connection between what is written and what is shown, while fusion-based systems use multiple encoders and then integrate the information, often leading to better interpretation of sarcastic or ironic content.

This study focuses on utilizing Large Language Models (LLMs) and Multimodal LLMs (MMLLMs) for detecting irony and sarcasm in social media posts. To evaluate our methods in a low-resource language, we also plan to create a Turkish multimodal dataset. For our experiments, we used two widely adopted datasets: MMSD 2.0 [3] and MuSE [4]. MMSD 2.0 contains around 24,000 ironic and non-ironic posts, while MuSE focuses on ironic examples accompanied by natural language explanations. The datasets are publicly available at https://github.com/JoeYing1019/MMSD2.0 and https://github.com/LCS2-IIITD/Multimodal-Sarcasm-Explanation-MuSE, respectively.

First, we performed zero-shot experiments by testing the Qwen-VL model with different prompts on both the MMSD 2.0 and MUSE datasets. We observed that the MMSD 2.0 dataset produced consistent results, while the MUSE dataset gave lower and more variable outcomes. Next, we used Qwen-VL to generate image captions for the images in the datasets, and we combined these captions with the original text. This combined input was then fed into a binary text classification model based on a transformer architecture, such as RoBERTa-base. We adapted our initial implementation from the codebase published by Ozturk et al. [5], which includes contributions from Research Assistant Fırat. Currently, we have tested the model on the entire MMSD 2.0 dataset and are experimenting with different configurations, including various learning rates and batch sizes.

Our approach differs from the original works in that we use Qwen-VL to generate image descriptions for each sample. We then combine these image captions with the original post text and pass the combined input to a RoBERTa-base binary classification model. We adapted our implementation from the work of Ozturk et al. [5], which aims to reduce stylistic bias in satirical datasets. The corresponding codebase is available at https://github.com/auotomaton/satireTR. Our current results are comparable to the baselines provided in the original papers, and we plan to improve performance further by experimenting with different models and training configurations.

This study aims to contribute to the field of multimodal irony detection by combining text and visual information using modern language models. Our work focuses on both experimenting with existing datasets and preparing resources for under-represented languages. Below are the main contributions of this research:

- We plan to create a new Turkish dataset for multimodal irony and sarcasm detection, including both text and image

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

data. This will support evaluation of existing methods in a low-resource language setting.
- We propose a method that uses Qwen-VL to generate image descriptions, which are then combined with the original text and passed to a RoBERTa-based model for binary classification.

REFERENCES

[1] Peiling Yi and Yuhan Xia, 2025. *Irony Detection, Reasoning and Understanding in Zero-shot Learning*, 14 pages. https://doi.org/10.48550/arXiv.2501.16884

[2] Abhilash Nandy and Yash Agarwal and Ashish Patwa and Millon Madhur Das and Aman Bansal and Ankit Raj and Pawan Goyal and Niloy Ganguly, 2024. *YesBut: A High-Quality Annotated Multimodal Dataset for Evaluating Satire Comprehension Capability of Vision-Language Models.* 18 pages. https://doi.org/10.48550/arXiv.2409.13592

[3] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. *MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System.* 10 pages. https://doi.org/10.48550/arXiv.2307.07135

[4] Desen Cai, Lu Wang, and Xiaojun Wan. 2021. "Nice perfume. How long did you marinate in it?" *Learning to Detect Sarcasm in Multimodal Social Media Posts.* In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 9 pages. https://doi.org/10.48550/arXiv.2112.04873

[5] Asli Umay Ozturk, Recep Firat Cekinel, and Pinar Karagoz. 2024. *Make Satire Boring Again: Reducing Stylistic Bias of Satirical Corpus by Utilizing Generative LLMs.* 17 pages. https://doi.org/10.48550/arXiv.2412.09247