Progress Report: Multimodal Irony/Sarcasm Detection

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

May 23, 2025

1. Goals and Purpose

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. Many people express themselves on social media using irony to make situations feel lighter or humorous, making irony detection an important topic in NLP.

While some approaches focus only on text or only on images, these single-modality methods are often insufficient. The full ironic meaning becomes clearer when both modalities are considered together. Therefore, multimodal methods that combine text and image are gaining popularity.

This study explores the use of vision-language models to generate image captions, which are then fused with the original post text. This combined input is classified using transformer-based models. Our experiments are conducted on MMSD2.0 and MuSE datasets, with the goal of improving irony detection by leveraging both modalities.

2. Data and Inputs

MMSD2.0 is a benchmark dataset of 24,000 social media posts labeled as sarcastic or non-sarcastic. It corrects issues found in earlier versions by removing spurious cues and re-annotating unreliable samples. We use the official train/validation/test split.

MuSE contains only sarcastic examples with paired image, caption, and explanation. It is used in our setup solely as a generalization test set.

Each image is captioned using Qwen-VL or InternVL. The generated description is concatenated with the original text and used as input to a binary classifier. Further details on the MMSD2.0 and MuSE datasets are available in [3, 4].

3. Method and Milestones

3.1. Overall Approach

We adopt a two-stage architecture: (1) caption generation from image using VL models; (2) fusion of caption and text, fed into RoBERTa or ModernBERT for binary classification. This modular design allows flexible pairing of components.

3.2. Progress So Far

We performed zero-shot evaluation on both datasets using Qwen-VL. Then, captions were generated for MMSD2.0 images using both Qwen-VL and InternVL. The fused inputs were used to train classifiers on MMSD2.0. Initial tests on MuSE show limited generalization, consistent with its single-class design. Our approach is conceptually similar to recent efforts in minimizing bias and improving multimodal sarcasm detection [5, 1].

4. Challenges

- Caption quality varies across vision-language models, affecting downstream performance.
- MuSE is single-class and geared toward explanation, limiting direct use in classification.
- Large models required significant GPU memory and time during captioning.

5. Preliminary Results

Fusing image captions with text improves accuracy on MMSD2.0 compared to text-only baselines. InternVL + ModernBERT achieved the best results. Zero-shot evaluation on MuSE was less effective, highlighting limitations of model transfer without negative samples. This aligns with previous findings that combining modalities leads to performance improvements [2].

Table 1: Best performing configuration per setting

Input	**Model**	**Acc**	**F1**	**Prec**	**Recall**
text-only	Qwen + Roberta	78.42	77.71	78.60	78.42
text-only	Qwen + ModernBERT	**84.72**	**84.59**	**85.21**	**84.72**
text-only	InternVL + Roberta	78.08	77.86	78.48	78.08
text+image	Qwen + Roberta	74.51	73.88	74.42	74.51
text+image	InternVL + Roberta	76.26	75.93	76.47	76.26
combined	Qwen + Roberta	82.07	81.88	82.42	82.07
combined	Qwen + ModernBERT	82.61	82.36	82.67	82.61
combined	InternVL + Roberta	81.82	81.62	82.14	81.82
Baseline	MMSD2.0	85.64	84.10	80.33	88.24

6. Next Steps

- Test our current approach on the MuSE dataset to evaluate generalization in a fully ironic, explanation-focused setting.
- Apply our pipeline to advanced multimodal models such as Gemini to compare performance with open-source alternatives.

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