TIB-VA at SemEval-2022 Task 5 A Multimodal Architecture for the Detection and Classification of Misogynous Memes

Preda Alexandru-Florin

Motivation

Online platforms are facing an increasing amount of misogynistic content. Memes are especially challenging because their meaning emerges from the interaction between images and text. Traditional text-only hate speech models often fail in these cases, making multimodal approaches necessary.

Task and Dataset

The challenge contains 2 tasks:

- Task-A: Binary classification of misogyny.
- Task-B: Multi-label classification of misogyny subtypes (stereotype, shaming, objectification, violence)

Dataset details- MAMI(Multimedia Automatic Misogyny Identification):

- Training set: 10,000 samples.
- Test set: 1,000 samples.

Splits	Task-A		Task-B				Total
	Misogynous	NOT	Shaming	Objectification	Violence	Stereotype	
Train	5000	5000	1274	2202	953	2810	10 000
Test	500	500	146	348	153	350	1000

Label: not misogynous

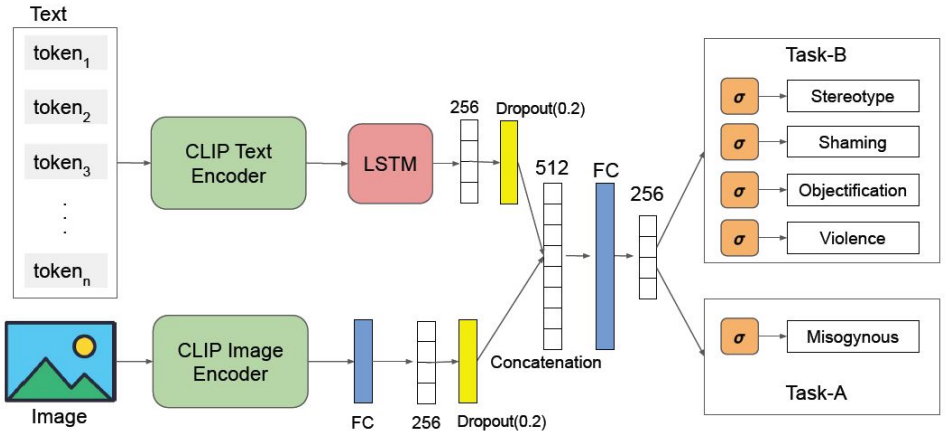
Label: not misogynous

Label: misogynous (stereotype, violence)

Label: misogynous (violence)

Model Architecture

- **Neural model that utilizes CLIP** for pre-trained multimodal features**:**
  - Separate encoders for text and image components.
  - LSTM layer for text context representation.
- **Combination** of text and image features via concatenation and fully connected layers.



Experimental Setup and results

Training Details:

- **Optimizer**: Adam.
- **Learning Rate**: 1e-4.(decreased by half every 5 epochs)
- **Batch Size**: 64.(max 20 epochs)
- **Validation Split**: 10% of the training data.

Evaluation Metrics: Macro F1 for Task-A and Weighted F1 for Task-B.

Overall Results

The model achieves a macro F1 score of 84.19 for binary misogyny detection in Task A. For Task B, which involves multi-label classification of misogyny types, the weighted F1 score is 61.60. These results are competitive with recent state-of-the-art systems.

My Results	Task-A	Task-B
ResNet50	0.83472	0.60974
ResNet504	0.84179	0.59460
vit14	**0.84190**	0.6160
vit32	0.83580	0.61527

Overall Results

The model performs best at detecting stereotypes and violence, while shaming is more difficult to classify accurately. This suggests that some misogyny types rely on more subtle contextual cues that are harder to capture, even with multimodal features.

Compared to other published systems from the SemEval-2022 challenge, this model achieves similar or better performance on binary classification. The results highlight the effectiveness of CLIP-based feature extraction and multimodal fusion strategies.

Conclusion and Future Work

This project demonstrates that multimodal learning is essential for effective misogyny detection in memes. Future improvements could include enhanced image analysis, OCR for stylized meme text, and extending the approach to other forms of hate speech beyond misogyny.