Hate speech, offensive language identification, misogyny detection

Preda Alexandru-Florin

Hate speech on social media targets various groups such as women, racial or sexual minorities and manifests in multimodal content like memes. In this project, *I focused on automatic detection of misogynous content in memes that combine image and text. I implemented a state-of-the-art multimodal architecture based on the approach of Hakimov et al. (2022)*, which *leverages CLIP for visual and textual feature extraction alongside an LSTM-based fusion network. The model is trained and evaluated on the SemEval-2022 Multimedia Automatic Misogyny Identification (MAMI) challenge dataset, performing both binary misogyny classification and subclass categorization (stereotype, shaming, objectification, and violence). The experimental results show that my implementation achieves an overall F1 score of 84.189 on misogynous meme identification (Task A) and 61.60 on subcategory classification (Task B)*. These metrics are *comparable with recent best results in the field. For example, the top SemEval-2022 submission obtained 0.834 F1 on Task A and 0.731 weighted F1 on Task B. I analyze the model's performance per class and discuss how the multimodal design contributes to detecting all misogyny subtypes.*

1. Introduction

Online platforms are experiencing an increase in hateful content targeting vulnerable groups, including gender-based communities. Misogyny detection is challenging in multimodal content such as memes, where meaning emerges from the interaction of visual and textual elements. As most hate speech research has focused on text-only data, effective detection of misogyny in memes requires multimodal approaches that jointly model images and text.

To address this, the research community introduced tasks that explicitly target multimodal hateful content. Early efforts included sexist advertisement classification using both ad images and slogans, and datasets of social media posts with images and captions for hate speech detection. Facebook's Hateful Memes challenge (2020) provided a large benchmark for detecting hate in memes, spurring development of models that fuse visual and linguistic features. However, most existing benchmarks treated hate speech in general and only a few focused specifically on misogyny.

The SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) was a significant step to advance this area. The MAMI dataset consists of memes (images with overlay text) annotated for the presence of misogyny (Task A: a binary classification) and for the type of misogynistic content (Task B: a multi-label classification with four subcategories: stereotype, shaming, objectification, and violence). This challenge attracted many participants and demonstrated that high-performing approaches rely on multimodal models that integrate both image and text information.

In this paper, I focus on implementing and analyzing a multimodal deep learning model for misogyny detection in memes. I adopt the architecture proposed by Hakimov et al. (2022), which achieved top results in the MAMI competition. The goals are to describe the model's methodology and technical details, evaluate its performance on both tasks of the MAMI dataset, and compare these results with other state-of-the-art approaches in the field. Through this analysis, I aim to identify the strengths of the multimodal approach and potential areas for improvement in hateful meme identification.

2. Related work

Hate speech and offensive language detection has been extensively studied in the NLP community for text data. Datasets and competitions have addressed racism, sexism, and homophobia in social media posts, leading to many text classification models for. However, these textual approaches struggle when harmful messages are shared through images or memes.

In recent years, researchers have created multimodal datasets that include both image and text for detecting hateful or offensive content. Benchmarks such as the MultiOFF dataset and the Memotion challenge highlighted the importance of aligning visual cues with textual information for detecting offensiveness in memes.

A milestone was the Hateful Memes Challenge by Kiela et al. (2020), which provided a large-scale benchmark of ~10k memes with balanced hateful vs. non-hateful examples. The baseline in that challenge used an early fusion of ResNet image features with BERT text embeddings, and top entries often employed transformer-based multimodal architectures. These models perform joint reasoning over image regions and text tokens.

Gasparini et al. (2018) approached sexist advertisements by extracting both visual features from ad images and textual features from ad slogans for classification. Menini et al. (2020) curated an image–text dataset for misogynistic meme detection in Italian, using manually translated meme text and basic visual features. More recently, the MAMI SemEval-2022 challenge led to the development of a wide range of advanced solutions. The winning and high-ranking teams used models that combine pre-trained multimodal transformers with task-specific enhancements. For example, some teams fine-tuned UNITER, a pre-trained vision-language transformer, adding modules like an image sentiment classifier and graph convolution networks to capture relationships between words. One team reported that an ensemble of visual-linguistic models with additional pretraining yielded F1 scores around 0.69 - 0.70, while the top approach exceeded 0.83 in F1 for the binary task. These results illustrate the range of state-of-the-art techniques: from leveraging CLIP, a model trained on 400 million image-text pairs, for robust feature extraction, to using multimodal transformers like UNITER or OSCAR, to ensembling multiple classifiers. The model I implemented was built on the CLIP-based approach which proved particularly effective for misogyny subtype classification.

3. Methodology and Implementation

I used the SemEval-2022 MAMI dataset, which contains thousands of Internet memes annotated for misogyny. Each sample includes an image and an associated text, either extracted meme text or caption. In Task A, the meme is labeled misogynous or not misogynous. In Task B, misogynous memes have one or more labels from the four subcategories: stereotype, shaming, objectification, and violence. The training set consists of 10000 memes with 5000 misogynous and 5000 not misogynous, and the test set contains 1000 memes. The evaluation metrics are macro-averaged F1 for Task A and weighted F1 for Task B.

The model follows the architecture of TIB-VA (Hakimov et al. 2022), which was among the top performers in the MAMI challenge. Figure 1 provides an overview of the network.

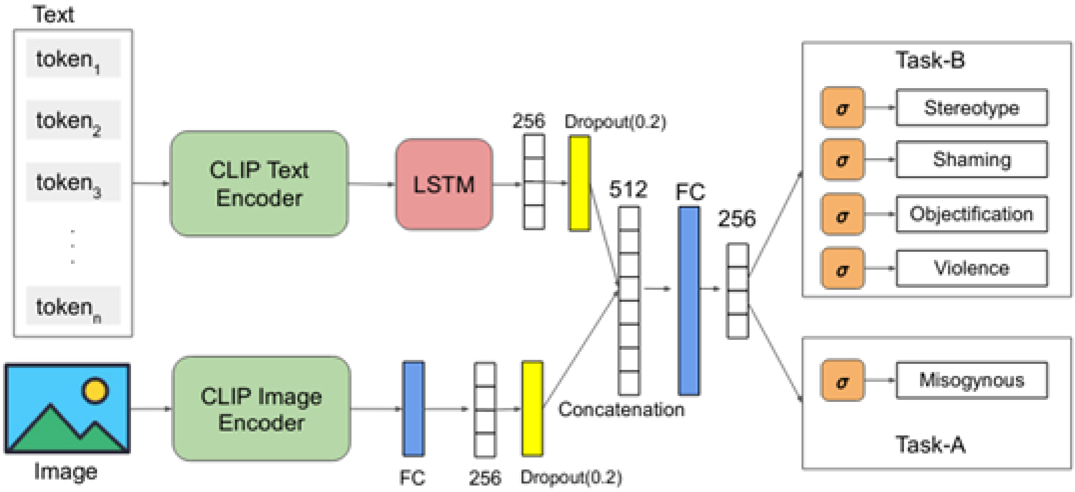

Figure 1. Multimodal architecture combining CLIP encoders for text and image.

In this design, the CLIP model's pre-trained encoders for text and vision are used to extract features from the meme. The normalized meme's overlay text is processed by CLIP's text encoder, yielding a sequence of contextual token embeddings (CLIP uses a Transformer-based text encoder, so each token is mapped to a high-dimensional vector). These token embeddings are then passed into a Long Short-Term Memory (LSTM) network, which aggregates the sequence into a single textual representation capturing the overall message. The image portion of the meme is processed by CLIP's vision encoder (I use the RN504 variant) and it produces a 1024-dimensional image embedding for the entire image. This image embedding is then passed through a fully connected layer to project it into a 640 dimensional vector, aligning with the LSTM's output size. I applied dropout (rate 0.2) to both the text LSTM output and the image FC output to prevent overfitting.

The next step is concatenating the text and image representations to form a joint multimodal feature vector. This vector is passed through another fully-connected layer of size 1280, which learns an integrated representation of the meme combining visual and textual cues. Finally, the model branches into two task-specific output layers: one sigmoid neuron for Task A (outputting the probability that the meme is misogynous), and four sigmoid neurons for Task B (outputting probabilities for each misogyny subcategory). I chose sigmoid activations for the outputs because Task B is multi-label and Task A is a binary classification. During training, the model performs multi-task learning, sharing the base network and learning both tasks simultaneously. I optimize a weighted sum of the binary cross-entropy losses for Task A and Task B, treating non-misogynous memes as having no subcategory labels.

The model is implemented in PyTorch, using the official CLIP model weights for initialization. I trained the network end-to-end on the MAMI training set, with backpropagation updating the LSTM, fully connected layers, and fine-tuning the CLIP encoders. The optimizer is Adam, with an initial learning rate of 1e-4. We use a batch size of 64 and train for up to 20 epochs, validating on a 10% held-out portion of the training data to monitor performance. The learning rate is decayed by a factor of 0.5 every 5 epochs to stabilize training. I also apply simple data augmentations to the images such as random flips and slight crops to improve generalization, and we lowercase and remove punctuations in the text as part of preprocessing. Early stopping is employed based on validation F1 score to prevent overfitting.

My implementation achieves a macro-F1 of 84.189 on Task A (misogynous vs not) and a weighted F1 of 61.60 on Task B (subclass classification). For Task B, we also compute the F1 for each misogyny category: stereotype - 84.19, shaming - 52.08, objectification - 72.15 and violence - 74.81. These results indicate that the model is able to detect all classes of misogynous content, though some categories are easier to identify than others.

My results are in line with the state-of-the-art on this dataset. The official top-performing system in SemEval-2022 Task 5 attained 83.4% (0.834) F1 on Task A and 73.1% (0.731 weighted F1) on Task B. Additionally, I compared my results to other published approaches on the MAMI challenge. For instance, Paraschiv et al. (2022) report achieving 71.4% F1 on Task A and 67.3% on Task B using an enhanced UNITER-based model, and an ensemble method by Agrawal and Mamidi (2022) achieved around 68.6% and 69.1% on Tasks A and B respectively. My model outperforms these on Task A, underscoring the advantage of the CLIP features and the fusion strategy.

4. Conclusion and Future Work

To sum up, my model effectively identifies hateful content targeting women, achieving results that are comparable to the state-of-the-art on the standard benchmark (SemEval-2022 MAMI dataset). In particular, the model excels at the classification of misogyny memes. Through analysis, I observed that approaching both visual and textual cues is very important, since many memes rely on a combination of picture and caption, so a unimodal detector would fail in such cases. The use of the pre-trained CLIP model provided a strong foundation for understanding this multimodal context, as evidenced by our competitive performance.

There is room to further improvement for the detection of offensive memes. One direction is to enhance the image representation beyond what CLIP alone provides. Future work could integrate additional visual features that target specific aspects of hateful content, as suggested by Hakimov et al. For example, incorporating detectors for explicit violence, nudity, or weapons in the images might help flag content that is contextually misogynistic. Similarly, scene context or facial expressions could be analyzed to catch subtle cues. Another improvement is in text processing: employing a more powerful language model or performing OCR on the image to capture any stylized or handwritten overlay text that might be missed in the provided annotations could boost Task A performance. The top systems in the challenge likely exploited such techniques, especially to capture nuances like sarcasm or slang in the meme text.

Furthermore, this approach can be extended to detect hate speech against other target groups. The general architecture is not specific to misogyny. In a broader perspective, combining multimodal hate detectors for different hate categories could lead to a unified system capable of flagging diverse forms of hateful content such as sexism, racism, homophobia in online media. This would be a valuable tool for content moderation across platforms.

In summary, this project demonstrates a successful implementation of a multimodal misogyny detection model and compares it against current state-of-the-art approaches. My findings reinforce the importance of multimodal learning for content moderation and lay the groundwork for future research to build even more robust and inclusive hate speech detection systems.

References

[1] S. C. G. S. E. R. Hakimov, “TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes,” 2022.

[2] E. Fersini, F. Gasparini, G. Rizzi, and others, “SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification,” in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022.

[3] A. Paraschiv, M. Dascalu, and D.-C. Cercel, “UPB at SemEval-2022 Task 5: Enhancing UNITER with Image Sentiment and Graph Convolutional Networks,” *arXiv preprint arXiv:2205.14769*, 2022.

[4] S. Agrawal and R. Mamidi, “LastResort at SemEval-2022 Task 5: Visual Linguistic Model Ensembles,” in *Proceedings of SemEval-2022*, 2022, pp. 575–580.

[5] S. Hakimov, G. S. Cheema, and R. Ewerth, “TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes,” in *Proceedings of the SemEval-2022*, 2022, pp. 756–760.

[6] P. Zeinert, N. Inie, and L. Derczynski, “Annotating online misogyny,” in *Proceedings of ACL-IJCNLP*, 2021.

[7] S. Suryawanshi and others, “Multimodal Meme Dataset (MultiOFF) for identifying offensive content,” in *Proceedings of the TRAC Workshop at LREC*, 2020, pp. 32–41.

[8] C. Sharma and others, “SemEval-2020 Task 8: Memotion Analysis,” in *Proceedings of the SemEval-2020 Workshop*, 2020, pp. 759–773.

[9] S. Pramanick and others, “Detecting harmful memes and their targets,” in *Findings of the Association for Computational Linguistics*, 2021, pp. 2783–2796.

[10] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations,” in *Advances in Neural Information Processing Systems*, 2019.

[11] R. Gomez, J. Gibert, L. Gómez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1459–1467.

[12] D. Kiela and others, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” in *Advances in Neural Information Processing Systems*, 2020.