Caption-augmented Multimodal Classification of Hateful Memes

Adem Kaya

Vrije Universteit, Amsterdam
a.kaya2@student.vu.nl

1 Introduction

Internet memes have emerged as a popular medium for delivering social commentary in the digital era. As they are growing more common, it is vital to recognize harmful content in order to preserve positivity in online communities. This has proven to be a challenging task, as the Facebook Hateful Memes (FHM) Challenge introduced in 2020 demonstrated that humans significantly outperform predictive models in the context of detecting hate speech in internet memes, with an accuracy of 84.7% [1].

The difficulty that arises is the intra-modal nature of memes; when considering solely the meme image or the meme text, it is not directly clear whether any harmful content is present, unless they are taken into account with relation to each other. While there has been a multitude of approaches [2–5] to this task, research is scarce on the efficacy of augmenting the hateful memes dataset with artificially generated captions. This study aims to fill that gap.

2 Methodology

The Vision-Language Model (VLM) known as BLIP-2 is a transformer model that achieves impressive results on vision-language tasks such as visual question answering and image captioning [6]. As such, the captioning capabilities of this model are used for the experiments, which consist of comparing two classification models: one trained on the original FHM dataset, and one trained on the BLIP-2-augmented FHM dataset.

The first step is to iterate through every meme in the dataset and use BLIP-2 to generate a corresponding caption, which is then added back to the dataset. Next, the features must be extracted from all the data. Again, state-of-the-art pre-trained models are used. For visual data, *CLIP* is used to extract visual features, as it is a model that excels at capturing visual information in images [7]. For textual data (captions and meme texts), I experiment with two models: BERT[8] and RoBERTa[9].

Finally, all feature vectors are concatenated into one vector to prepare for the training process of the classification model. The classification model is a neural network that outputs a binary classification to predict the label (hateful/non-hateful). Refer to Figure 1 for a visual representation of the process.

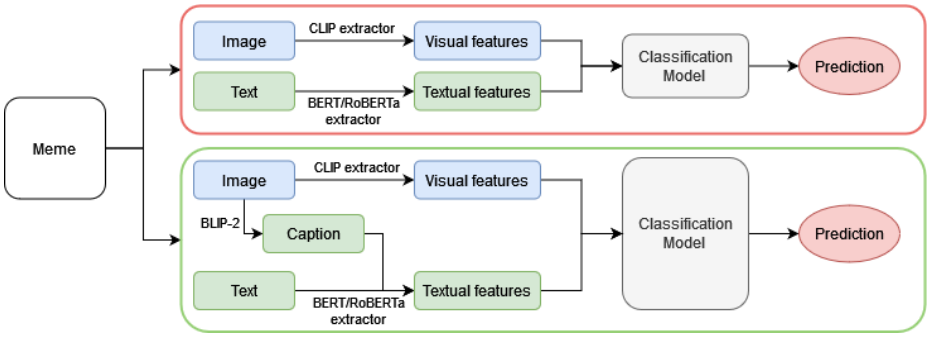

Fig. 1. The two approaches taken: the baseline classification model trained on features extracted from the original FHM dataset, and my method trained on the features of the caption-augmented set.

3 Results and Discussion

Refer to Table 1 for the results of the experiment. The top three rows are baseline metrics from the FHM Challenge paper. It can be observed that the metrics of the models trained on the caption-augmented sets do not suggest any improvement in performance. While the baseline (i.e. trained on the unaltered dataset) models managed to achieve AUROC scores of 0.711 and 0.698, respectively, there were no significant changes observed in the experimental models.

Table 1. Classification results for the baseline and experimental models.

Model	Validation Accuracy	Test Accuracy	Test AUROC
FHM: Human		0.847	
FHM: ViLBERT CC	0.661	0.659	0.745
FHM: Visual BERT COCO	0.659	**0.695**	**0.754**
Original (BERT+CLIP)	**0.744**	0.614	0.711
Original (RoBERTa+CLIP)	0.730	0.632	0.698
Captioned (BERT+CLIP)	0.732	0.672	0.716
Captioned (RoBERTa+CLIP)	0.740	0.596	0.683

This indicates potential issues with the experimental setup. One main point is whether the visual and textual features align semantically, since CLIP and BERT/RoBERTa have distinct embedding spaces. This may lead to ineffective learning. Another factor is the quality of the captions; if BLIP-2 can not meaningfully capture any nuanced meanings of jokes or references, it only serves to add a layer of unneccesary complexity. In the paper, a more thorough analysis is performed on the caption quality, and different ideas are explored about possible future work to address these issues.

Acknowledgement

I thank my supervisor Dr. Filip Ilievski for their continued support throughout this research. Their expertise in this field was invaluable in shaping the outcomes of this work.

References

1. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. (2021)
2. Cao, R., Ka-Wei Lee, R., Chong, W., Jiang, J.: Prompting for Multimodal Hateful Meme Classification. (2023)
3. Junhui, J., Wei, R., Usman, N.: Identifying Creative Harmful Memes via Prompt based Approach. Proceedings of the ACM Web Conference 2023, 3868–3872 (2023)
4. Grasso, B., La Gatta, V., Moscato, V., Sperlì, G.: Kermit: Knowledge Empowered model in harmful meme detection. Information Fusion 106, 102269 (2024)
5. Shan Hee, M., Chong, W., Ka-Wei Lee, R.: Decoding the Underlying Meaning of Multimodal Hateful Memes. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 5995–6003 (2023)
6. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. (2023)
7. Radford, A., Wook Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. (2021)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional
9. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019)