Caption-augmented Multimodal Classification of Hateful Memes

Vrije Universiteit, Amsterdam, the Netherlands
a.kaya2@student.vu.nl

Abstract. Hateful Meme Classification is a task that can be challenging due to its multimodal nature which contains images and text. In this research, I specifically investigate whether this task benefits from using captions generated by the Vision-Language Model (VLM) know as BLIP-2. While there have been many models developed for this task which also use captioning in their architecture pipeline, the research on the benefits of solely adding captions is scarce. I approach this task by generating captions for the images, extracting the visual features from the images, and extracting the textual features from both the captions and meme texts, which are finally appended to each other. The evaluation shows no improvement of the models after training it on caption-augmented data; potential reasons for the lack of performance changes and ideas to investigate these issues are thoroughly discussed in the Analysis and Discussion sections.

Keywords: Hateful Memes · Multimodal Classification · Dataset Augmentation · Image-to-text Captioning.

Disclaimer: This paper contains examples of hateful memes with discriminatory content that may be disturbing to some readers.

1 Introduction

In today's digital age, memes have become a prominent medium for carrying out social commentary in the form of humour. As memes are rapidly growing more and more common, it becomes necessary to deploy systems that can process memes automatically. As such, being able to accurately detect harmful content within memes is crucial for maintaining positivity in online communities.

In 2020, Facebook introduced the Hateful Memes Challenge (HMC) set, together with an article evaluating human performance against various models in the context of detecting hate speech [1]. According to the results, humans significantly outperform state-of-the-art unimodal and multimodal models with an accuracy of 84.7%, indicating that such models generalize poorly on memes. Due to the multimodal nature of memes, it becomes challenging to develop a model that can comprehend not only the images and their texts, but rather the interaction between the two, since memes generally have a direct relationship

between the images and texts. The prominence of sarcasm, irony, and satire as comedy tropes as well as the reliance on inside jokes and cultural references contribute to an added level of difficulty. There have been various attempt to tackle this problem, such as using a prompt-based approach [2, 3], or enriching models with external knowledge [4, 5]. While these approaches are able to achieve high-performing models, I focus on the scarcity of research on the topic of augmenting the hateful memes dataset with captions, and determining whether the classification task benefits from these captions.

In this work, I aim to answer the research question:

> *Does augmenting the Facebook Hateful Memes dataset with captions generated by a state-of-the-art Vision-Language Model (VLM) improve the performance of a hatefulness classification model?*

The VLM that is used in this research is know as BLIP-2, which excels at multimodal tasks such as visual question answering and image captioning [6]. Specifically, the objective is to determine whether the context provided by the captions are relevant and useful, and thus, improve the performance of hatefulness detection models. The experiment is performed by training two types of classification models: one trained on the original FHM dataset with no augmentations, and one trained on the caption-augmented dataset.

See Figure 1 for three examples of BLIP-2's captions on one non-hateful meme and two hateful memes. When considering the meme in the middle, its corresponding text "*would you look at that, this dryer comes with a free dishwasher*" is not inherently harmful, yet if the image is also taken into account, it becomes evident that the meme is referring to a woman as "dishwasher", perpetuating a harmful stereotype. This suggests an intra-modal relationship (i.e. the image and text are related), which is common in memes. Furthermore, the corresponding caption, "*a woman is looking out of a dishwasher*", although containing a slight inaccuracy (the woman is looking out of a dryer), shows that BLIP-2 is able to successfully identify that a woman is present in the meme, so I expect the classification model that is trained on captions is able to learn more of such details.

2 Related Work

2.1 Hateful Meme Classification

Thanks in large part to Facebook's Hateful Memes Challenge, the classification of hateful memes has gained more prominence and become more public as a multimodal task. In the original article [1], some performance benchmarks were provided, where multimodal methods such as ViLBERT [7] and VisualBERT [8] were among the top performing models, indicating that multimodal methods, especially transformer architectures, are suitable for this task.

The winning solution [9] of the challenge describes using an ensemble of four different transformer architectures to achieve an AUROC score of 0.845,

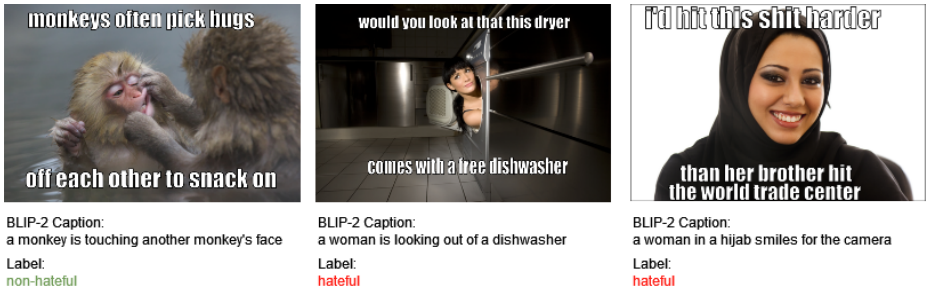

Fig. 1. Examples of BLIP-2's captioning capabilities on three random memes sampled from the HMC dataset. The left meme is considered non-hateful while the others are hateful.

namely VL-BERT [10], UNITER [11], VILLA [12], and ERNIE-VIL [13]. These models are all Visual-Linguistic (VL) in nature, meaning they can learn from both images and text simultaneously.

Prompting-based approaches also emerged as succesful methods of hateful meme classification. Cao et al. [2] describe how PromptHate, their proposed model which uses simple prompts combined with the pre-trained RoBERTa [14] language model, yields an AUROC score of 90.96. Likewise, Ji et al. [3] demonstrate how using a prompt-based language model outperforms state-of-the-art methods. However, the benefit of using captions does not immediately become clear from these methods.

2.2 Data Enrichment and Augmentation

Beyond the development of models, there have been methods to enrich the dataset with additional data. KnowMeme [15] attempts to build a knowledge graph to leverage external facts to aid in effectively detecting hate speech. The study highlights the importance of knowledge-informed decision making in the context of harmful meme detection. A recent study proposes an improved solution: KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection) [4], which builds a network of relevant knowledge obtained from ConceptNet [28] (a semantic network represented by a knowledge graph that connects words with labeled edges), while also considering relationships between entities within the knowledge graph.

HatReD [5] (Hateful meme with Reasons Dataset) focuses on the explainability aspect of hatefulness detection. By annotating Facebook's Hateful Memes dataset with underlying reasons, they propose a novel dataset that aims to address the research gap between meme explainability and hatefulness detection. A task at WOAH 5 [16] also builds on top of the Hateful Memes dataset by annotating it with the protected categories (i.e. the targeted demographics), and the attack types (e.g. inciting violence, dehumanizing).

Similarly, augmentation (i.e. artificial generation of new data) is used to aid in classification tasks. Li et al. [17] demonstrate how training on datasets generated from augmentation leads to an increase in AUROC score. Another method by Li et al. [18] focuses specifically on a low-resource solution in terms of computation, as existing research relies on complex solutions that require substantial computational resources. The idea consists of reducing the multimodal task into a single-modal (i.e. text) task, by generating image-to-text captions and appending them to the meme texts while maintaining high accuracy. Pro-Cap [19] deploys a comparable solution by generating captions, but combines it with a visual question answering (VQA) method to obtain more detailed information.

In this study, I augment the dataset with artificially generated captions. While [18] had a similar idea, this approach differs in the fact that I maintain the multimodal aspect of the task. Similarly, as opposed to [19], which uses VQA, this approach will only use captioning. The task will remain in its simplest form, and as such, the impact of the captions on multimodal classification can be measured.

3 Method

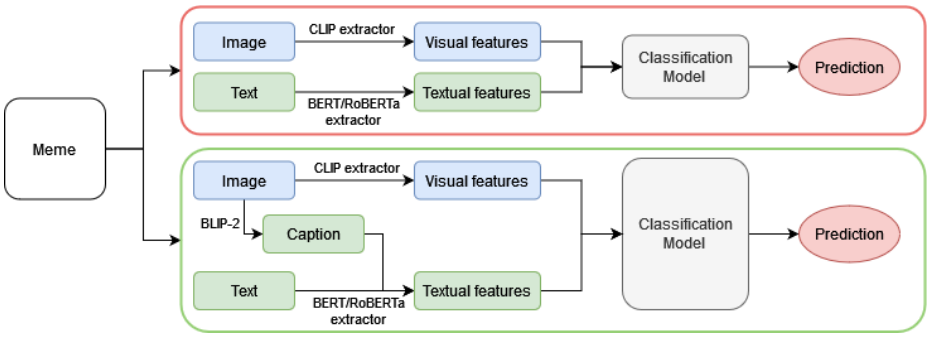

Fig. 2. The two approaches taken, classification using only image and text, as well as the experimental approach of adding captions.

The experimental method taken for this research consists of building two types of classification models; one that is trained on the Hateful Memes dataset with no additional augmentation, and one that is trained on the same dataset augmented with captions generated by BLIP-2. This is further illustrated in Figure 2.

3.1 Caption Generation

The process of caption generation is performed by simply iterating through every image in the dataset, and running $BLIP2$ [6]. Specifically, the model used is the OPT-2.7b model that is finetuned on COCO [29], which is a large dataset for object detection and captioning. These captions are then added back to the original dataset, preparing it for the feature extraction process.

3.2 Feature Extraction

For the training process, the features must be extracted from the multimodal data: visual features from the images, and textual features from the texts (both the captions and the meme texts). I decided to use pre-trained models for both feature extraction processes, as described below.

To extract visual features from the image data, I use the CLIP model [20]. CLIP is a state-of-the-art visual-language model that excels at visual tasks. It is trained on 400 million (image, text) pairs, where it learns to associate natural language description with visual concepts. Since the model generalizes well on data in various domains, and has also been used prior in Hateful Meme Classification tasks [22], it seems like a reasonable choice to capture visual information from the meme images.

For the textual feature extraction process, I experiment with two models: $BERT_{base}$ [21] and $RoBERTa_{large}$ [14].

BERT is trained on Wikipedia (2.5B words) as well as Google's BooksCorpus (800M words), and is used for a wide variety of natural language processing tasks. RoBERTa is a similar model that builds upon BERT and improves it to the point that it outperforms it in various natural language processing (NLP) tasks [14]. Since these models are bidirectional, they excel at capturing context in texts.

Consequently, the embeddings extracted from CLIP and BERT/RoBERTa will be concatenated and used by the classification model as training data.

3.3 Classification Process

The visual features extracted from the images, and the textual features extracted from the texts and captions are used to train two types of models. The architecture of these models is displayed in Table 1. The distinction between these two types of models is the text input layer. For the original model, this is simply the size of the textual features, but for the captioned model, this becomes $textfeaturesize + captionfeaturesize$.

Ultimately, there are four training processes for the classification model:

1. Original dataset, using BERT.
2. Original dataset, using RoBERTa.
3. Caption-augmented dataset, using BERT.
4. Caption-augmented dataset, using RoBERTa.

Table 1. Classification model layers and parameters for the models.

Layer	Shape (in, out)	Activation
Image input layer	(Visual features size, 256)	ReLU
Text input layer	(Textual features size, 512)	ReLU
Concatenated layer	(256 + 512, 128)	ReLU
Dense layer	(128, 64)	ReLU
Output layer	(64, 1)	Sigmoid

3.4 Pipeline

The full pipeline that is employed becomes:

1. Generate image-to-text captions using BLIP-2.
2. Extract visual features using CLIP.
3. Extract textual features using BERT/RoBERTa.
   (a) Append the caption's textual features to the meme text's features
4. Concatenate the visual and textual features.
5. Train a binary classification neural network to predict the label (hateful/non-hateful).

4 Evaluation

A key step to determining the performance of the various models is the evaluation. Since the original Facebook Hateful Memes challenge paper provided a baseline evaluation of humans in terms of accuracy and AUROC score [1], I perform my evaluation in the same manner to maintain consistency.

The evaluation is performed in two parts: during the training process, a validation partition is used to calculate the validation accuracy. The validation set is partitioned from the training dataset with a size of 20%. Furthermore, a test partition containing 500 rows is used to determine the remaining metrics, which contains 254 non-hateful memes and 246 hateful memes. However, one thing to note is that the train partition consists of 5,493 non-hateful memes and 3,007 hateful memes, which means there is a heavy class imbalance.

Computation Environment To access the BERT and RoBERTa models, the Transformers module from Huggingface [26] was used, which is a library that provides many pre-trained transformer architectures under an API. The CLIP model followed the implementation provided by OpenAI [20]. Finally, the BLIP-2 captioning model was implemented using the library provided in LAVIS (A Library for Language-Vision Intelligence) [27].

Consequently, the classification models are trained in Python Notebooks on a desktop computer with the following specifications:

- Processor: AMD Ryzen 5 5600
- GPU: NVIDIA RTX 3060 TI

– Memory: 32 GB RAM
– Storage: 512 GB SSD

Using the following configuration details for the training process:

– Learning Rate: 0.001
– Batch Size: 32
– Epochs: 30

5 Results

The results of the experiment can be seen in table 2. The observed results suggest potential issues with the experimental setup; it is evident that there are no noticeable differences between the test accuracy and test AUROC score of each model, which are approximately 0.64 and 0.71, respectively. Thus, this section is dedicated to analyzing the potential complications that have emerged during the experiment.

Table 2. Classification results for all models. A few baseline metrics from the original Hateful Memes Challenge paper are displayed on top.

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

A primary concern is whether the features extracted from the pre-trained models align semantically. Both CLIP and BERT/RoBERTa have distinct spaces for embeddings [14, 20, 21]. Although CLIP is a multimodal model trained on (image, text) pairs, I extracted solely visual features using CLIP. Hence, the visual embeddings may not align completely with BERT/RoBERTa's textual embeddings in the semantic space. In such cases of misalignment, the model may not be learning the semantics effectively, and as a result, the augmentation of the data with BLIP-2 captions does not provide added understanding of hateful memes.

Another factor that could account for the lack of improvement is simply the quality of the captions. While BLIP-2 excels at visually describing images as seen in Figure 2, these descriptions may not provide any additional meaningful information in the context of memes. It is also unlikely that BLIP-2 is capable of capturing any nuanced meanings of inside jokes or cultural references. Then, the possibility arises that the captions are in fact counter-productive by adding unnecessary complexity, which can confuse the model during training.

Caption Quality Analysis I perform a more in-depth analysis of the predictions to determine whether the captions provide utility to the models. I select 30 samples from the caption-augmented test partition of the dataset to manually assess the caption quality in terms of usefulness and relevance using a 3-point Likert scale. Then, these assessments can be contrasted against the caption's utility to the model to obtain an objective evaluation. The result of the manual assessment is presented in table 3.

Table 3. Caption quality assessment results (n=30).

Not useful at all	**Moderately useful**	**Very useful**
8 (26.67%)	12 (40%)	10 (33.33%)

During this analysis, it became evident that the captions which were least useful were from memes in which the entity or person is of importance, which BLIP-2 was unable to recognize. Figure 3 shows an example of a caption I assessed to be not useful at all. The reasoning for this is the fact that the meme implies that Hillary Clinton tends to orchestrate suicides on her political dissidents. Since BLIP-2 is unable to recognize that this is Hillary Clinton, the classification model would not be able to learn to make this connection. However, although I would personally consider this meme to be hateful, it is labeled as not hateful in the dataset, and the predictions of all models are still correct for this case. This highlights an issue in the context of hateful meme detection: the hatefulness of a meme is debatable and open to interpretation.

Other prominent examples of memes where the captions are not useful are memes containing a collage of images, which seemingly confuses BLIP-2, as it is trained to describe an image as a whole, as opposed to multiple images in one. This is likely counter-productive to the classification model. This is shown in figure 4; in this particular case, the predictions on both captioned models have changed from a correct classification to an incorrect classification.

Furthermore, figure 5 shows an example where the caption was assessed to be very useful, which consequently improved the captioned BERT model's performance; however, the RoBERTa model's prediction stayed the same. The other results of the 30 analyzed samples are displayed in Table 4. These results suggest that the captions mostly make no impact, as in both models the majority of the predictions stayed the same. Moreover, the RoBERTa model seems to obtain worsened predictions after captioning, more often than the BERT models; however, given that thirty is not a completely representative sample size, a more thorough investigation would be required to determine the specific impact of captions on the different models.

Table 4. Predictions of the manually assessed samples before and after captioning.

Model	No change (correct)	No change (incorrect)	Improved	Worsened
BERT	17 (56.67%)	5 (16.67%)	6 (20%)	2 (6.67%)
RoBERTa	15 (50%)	7 (23.33%)	2 (6.67%)	6 (20%)

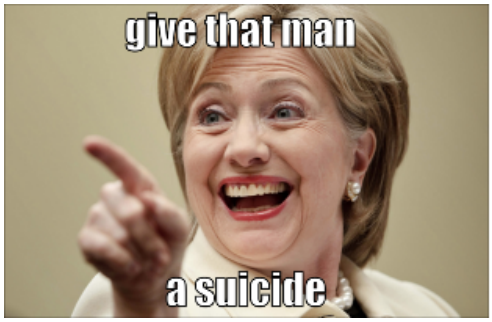

Fig. 3. Example of a caption that is deemed not useful at all.

Fig. 4. Second example of a caption that is deemed not useful at all, causing the predictions to become incorrect.

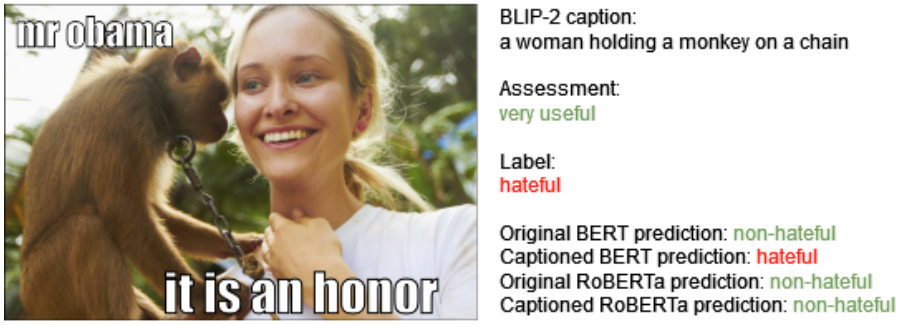

Fig. 5. Example of a caption that is deemed very useful, causing an improvement in prediction.

6 Discussion

Since the results do not contain the expected content, they can be considered a null result. This section will more thoroughly discuss the interpretation of the results and methods to further investigate the issues mentioned in the Analysis subsection.

Aligning Textual and Visual Representations To address the potential issue of misalignment of visual and textual embeddings, the CLIP and BERT/RoBERTa models could be fine-tuned specific to the task at hand to enable better learning of semantic representations. The paper of a proposed model named Hate-CLIPper [22] that uses CLIP for hateful meme classification mentions the use of *projection layers* to achieve better alignment between the image and text spaces. Moreover, the article suggests that modeling the interactions between image and text features using a *feature interaction matrix* (FIM) is better than simply concatenating the features, which is what I did in this experiment. The concatenation of features ostensibly fails to properly capture the relationships between the features. Another paper by Lee et al. [23] focusing on hateful meme classification specifies also using a BERT encoder for their classification; however, they initialize it by fine-tuning it to their task. Additionally, to better model the interaction between the visual and textual modalities of the task, they mention applying an attention mechanism proposed by Vaswani et al. [24]

Multimodal Methods Furthermore, instead of using CLIP and BERT/RoBERTa separately, an alternative effort is to employ a multimodal model such as VisualBERT [8] or ViLBERT [7]. The multimodal nature of these models ensures that the visual and textual features are embedded in the same space. The third place [25] of the Facebook Hateful Memes challenge describes using VisualBERT, and achieving a test accuracy of 0.765 and an AUROC score of 0.811.

Caption Quality To determine the presence of any issues regarding BLIP-2 captioning, there could be a qualitative research performed to observe the quality of the captions in the context of this particular task. Furthermore, BLIP-2's visual question answering (VQA) capabilities are potentially better to use for this task, as they can be utilized to elicit demographic details about the image. This is the approach taken in the Pro-Cap paper mentioned in section 2 [19]. For instance, they inquire specifically about the race of the person in the image. Doing this for other common targets of hateful memes, they are able to collect a considerable quantity of contextual information, which likely aids the training process.

7 Future Work

Possible future work could contain many different adaptations and experiments with regards to the task explored in this paper. The primary concern is the adoption of different methods and pipelines. Furthermore, different choices of datasets are also opportunities for exploration. These are the ideas that could be tested in future works regarding caption-augmented hateful meme classification:

1. Using different versions of the Facebook Hateful Memes dataset. Specifically, the datasets described in section 2.2 could be experimented with (HaTReD [5], and the task from WOAH [16]).
2. Adopting different methods to improve, or wholly change, the feature extraction process. Fine-tuning CLIP and BERT/RoBERTa on the data may lead to better results, as well as using specific techniques to apply attention mechanisms to better represent the joint features between images and text.
3. Employing alternative pre-trained models for classification, such as VisualBERT or ViLBERT. This manner eliminates the necessity to use two distinct models for the two modalities of the task, removing a layer of complexity.
4. Performing a comparative study on augmented hateful meme classification between BLIP-2's captions and VQA capabilities to see which technique is more suited for the task. Using VQA may provide more context for the model to work with.

8 Conclusion

This study set out to answer the following question:

> *Does augmenting the Facebook Hateful Memes dataset with BLIP-2 generated captions improve the performance of a hatefulness classification model?*

While the model was able to achieve an AUROC score of approximately 0.71, which is significantly better than the baseline performance of 0.5, there were no improvements in both the accuracy and AUROC score metrics when comparing

the original dataset with the caption-augmented dataset. Thus, the results can be concluded to be a null result.

During the Analysis and the Discussion, some potential issues that could have contributed to the results were highlighted and discussed. The lack of any intra-modal mechanism to represent correlations between visual and textual embeddings, and not using BLIP-2's capabilities to their fullest may have contributed to the null result.

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. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. (2019)
8. Harold Li, L., Yatskar, M., Yin, D., Hsieh, C., Chang, K.: VisualBERT: A Simple and Performant Baseline for Vision and Language. (2019)
9. Zhu, R.: Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. (2020)
10. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of Generic Visual-Linguistic Representations. (2020)
11. Chen, Y., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: UNiversal Image-TExt Representation Learning. (2020)
12. Gan, Z., Chen, Y., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-Scale Adversarial Training for Vision-and-Language Representation Learning. (2020)
13. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. (2021)
14. 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)
15. Shang, L., Youn, C., Zha, Y., Zhang, Y., Wang, D.: KnowMeme: A Knowledge-enriched Graph Neural Network Solution to Offensive Meme Detection. 2021 IEEE 17th International Conference on eScience (eScience), 186–195. (2021)
16. Lambert, M., Shaolian, N., Douwe, K., Vinodkumar, P., Bertie, V., Zeerak, W.: Findings of the WOAH 5 Shared Task on Fine Grained Hateful Memes Detection. Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 201–206. (2021)

17. Li, Y., Zhang, Z., Huang, H.: Enhance Multimodal Model Performance with Data Augmentation: Facebook Hateful Meme Challenge Solution. (2021)
18. Li, Y., Chan, J., Peko, G., Sundaram, D.: Towards Resource Inequities in Catching the "Dark Side" of Social Media: a Hateful Meme Classification Framework for Low-resource Scenarios. Proceedings of the 57th Hawaii International Conference on System Sciences, 7215–7224. (2024)
19. Cao, R., Shan Hee, M., Kuek, A., Chong, W., Ka-Wei Lee, R., Jiang, J.: Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection. (2023)
20. 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)
21. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional
22. Kumar, G. K., Nandakumar, K.: Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. (2022)
23. Lee, R. K., Cao, R., Fan, Z., Jiang, J., Chong, W.: Disentangling Hate in Online Memes. Proceedings of the 29th ACM International Conference on Multimedia. (2021)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. (2023)
25. Velioglu, R., Rose, J.: Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. (2020)
26. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A. M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. (2020)
27. Li, D., Li, J., Le, H., Wang, G., Savarese, S., Hoi, S. C. H.: LAVIS: A One-stop Library for Language-Vision Intelligence. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 31–41 (2023)
28. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. (2018)
29. Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., Dollár, P.: Microsoft COCO: Common Objects in Context. (2015)