AAST-NLP at Multimodal Hate Speech Event Detection 2024 : A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models.

Ahmed El-Sayed and **Omar Nasr**
Arab Academy For Science and Technology
{ahmedelsayedhabashy,omarnasr5206}@gmail.com

Abstract

With the rapid rise of social media platforms, communities have been able to share their passions and interests with the world much more conveniently. This, in turn, has led to individuals being able to spread hateful messages through the use of memes. The classification of such materials requires not only looking at the individual images but also considering the associated text in tandem. Looking at the images or the text separately does not provide the full context. In this paper, we describe our approach to hateful meme classification for the Multimodal Hate Speech Shared Task at CASE 2024. We utilized the same approach in the two subtasks, which involved a classification model based on text and image features obtained using Contrastive Language-Image Pre-training (CLIP) in addition to utilizing BERT-Based models. We then utilize predictions created by both models in an ensemble approach. This approach ranked second in both subtasks, respectively.

1 Introduction

Social media has become the biggest form of communication in recent years. However, with this rise comes an increase in the usage of hate speech to spread hostile and hateful messages. The effects of hate speech have been very apparent in recent years and have been demonstrated in multiple studies (Parihar et al., 2021). Some malicious entities have even been shown to use memes to create such hateful content. While these memes might seem humorous in nature, studies show that this use of humor to spread hateful messages creates hostile perceptions within the audience (Schmid, 2023). The use of machine learning and AI to combat this problem and classify these memes has been on the rise lately, with the collection of large amounts of data and the creation of datasets to support these tasks (Kiela et al., 2021). The use of such hateful attacks has been widespread and particularly evident in the Russia-Ukraine conflict, where both parties engaged in masquerading these attacks as memes. The Multimodal Hate Classification shared task at CASE 2024 (Thapa et al., 2024) focused on tackling this problem by providing a multimodal dataset primarily focused on this conflict (Bhandari et al., 2023). The rest of this paper is dedicated to our approach in the two subtasks provided in this shared task where we utilized CLIP (Radford et al., 2021) in conjunction with concatenation and a classification head to achieve second place on both subtasks. The following sections of the paper will include a related work section, a section describing the dataset, a section describing the system proposed, a discussion section and a conclusion.

2 Related Work

Research has extensively explored the application of AI in hate speech detection. However, fewer studies have delved into the use of multimodal data for classifying memes in these contexts. One notable study is (Pramanick et al., 2021), where they employed four different models for feature extraction, including CLIP, VGG-19 (Simonyan and Zisserman, 2015), and DistilBERT (Sanh et al., 2019), complemented by a CMAF fusion layer at the end. Another innovative approach was introduced by (Kumar and Nandakumar, 2022), presenting the HateClipper architecture. They utilized CLIP for feature extraction and implemented various fusion methods, such as alignment, concatenation, and cross fusion, resulting in promising outcomes. Additional methodologies were elucidated in (Cao et al., 2023), where researchers leveraged prompts and language models for classification. CASE 2023 featured a similar shared task (Thapa et al., 2023), with (Sahin et al., 2023) employing an ensemble of syntactical feature outputs passed into XGBOOST (Chen and Guestrin, 2016), coupled with encoder outputs, to achieve their noteworthy results.In recent times, researchers have presented datasets aimed at addressing this issue

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 139–144

(Thapa et al., 2022; Bhandari et al., 2023). These datasets mark a significant advancement, providing researchers with valuable resources to more effectively confront the problem and explore various architectures. One Interesting approach is the one proposed by (Yang et al., 2022) which incorporated a multimodal backbone with three additional modules semantic adaptation module, definition adaptation module and domain adaptation module which boosted the performance significantly.

3 Dataset & Task

The Multimodal Hate Speech Event Detection challenge at CASE 2024 (Thapa et al., 2024)[1] encompasses two specific subtasks: Subtask A and B. The dataset makes use of the CrisisHateMM dataset (Bhandari et al., 2023) which is a collection of Text-Embedded Images of Directed and Undirected Hate Speech from Russia-Ukraine Conflict. The following subsections expand on each subtask highlighting the data distribution of each label. For the test labels, these labels remain undisclosed and are reserved for assessing the ultimate prediction performance, influencing the leaderboard rankings at the conclusion of the shared task.

3.1 Subtask A: Hate Speech Detection

The first subtask is a binary classification problem where tweets given are classified into two distinct classes: "Hate Speech" and "No Hate Speech". Table 1 illustrates the data distribution for the different classes within the dataset.

	Training	Validation
No Hate	1658	200
Hate	1942	243
Overall	3600	443

Table 1: Subtask A's Dataset Distribution.

3.2 Subtask B: Target Detection

The second subtask is a multiclass classification problem where tweets given are classified into three distinct classes: "Individual", "Organization", and "Community". Table 2 illustrates the data distribution for the different classes within the dataset.

	Training	Validation
Individual	823	102
Organization	784	40
Community	335	102
Overall	1942	244

Table 2: Subtask B's Dataset Distribution.

3.3 Textual Data Extraction

The Google Vision API[2] was employed to extract textual information embedded in images. While the API demonstrates commendable accuracy and delivers high-quality results, its utilization is financially challenging for numerous researchers. This situation prompts the exploration of alternative tools such as various open-source Python packages or the creation of a comparable tool that maintains high quality but at a significantly lower cost.

3.4 Textual Data Preprocessing

Prior to being fed into the model, the text undergoes a rigorous preprocessing stage aimed at addressing various challenges related to the nature of social media data, where texts contain relatively high noise. This noise, if not properly handled, has the potential to adversely impact our classifier's performance. Therefore, the preprocessing stage is crucial in mitigating such adverse effects and ensuring the robustness of the model against the inherent noise in social media texts.

- Removal of punctuation as many tweets contained .
- Applying PySpellChecker[3] to check for misspelled words and correct them.
- Removal of hyperlinks as they did not meaning needed for our classification process.
- Removal of hashtags.

3.5 Visual Data Preprocessing

No preprocessing was applied to the images except for resizing them to dimensions of 224 x 224 pixels.

4 Methodology

In the following subsections, we will expand on the proposed models for each subtask, highlighting the reasoning behind each.

[1] https://codalab.lisn.upsaclay.fr/competitions/16203

[2] https://cloud.google.com/vision

[3] https://pypi.org/project/pyspellchecker/

4.1 Language Models

Language Models were found to achieve state of the art performance on many tasks including ones related to hate speech detection. After examining existing literature on multimodal hate speech detection, we discovered that relying solely on textual features yielded commendable results, approaching those achieved by approaches incorporating multiple modalities (Singh et al., 2023; Aziz et al., 2023). Consequently, we opted to conduct experiments employing several pretrained language models, with a primary focus on HateBERT (Caselli et al., 2021), RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2020).

4.2 Vision Models

After a thorough review of past literature, including research findings from last year's competition (Thapa et al., 2023) and various other sources, we have chosen not to investigate vision models on their own, as their performance on comparable tasks has been relatively subpar. Instead, our strategy entails conducting experiments with them as feature extractors within our multimodal framework. In the multimodal approach we adopted, we opted to employ ViT (Dosovitskiy et al., 2021) and Swin (Liu et al., 2021) as feature extractors.

4.3 Multimodal Approach

The multimodal approach comprises two main models that will be elucidated in the subsequent subsections.

4.3.1 Pairing Models

The initial approach aimed to harness the combined capabilities of vision and language models as this approach proved to beneficial in similar settings(Chen and Pan, 2022; Das et al., 2020). In the experiments, both language and vision models were employed as feature extractors without undergoing model finetuning. Subsequently, as part of the training procedure, both models were finetuned. The finetuned models yielded slightly higher results compared to the alternative. Throughout our exterminations, we experimented with pairing a number of models yet only 2 of them were used for submitting results through the official contest page as they mostly produced bad results. One important aspect to mention is the fact that our Swin + HateBERT model used the pretrained model weights without any further fintuning whilst the ViT + HateBERT model was fully fine tuned on the chosen dataset.

4.3.2 CLIP

State-of-the-art performances on numerous subtasks have been achieved by CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021). High results on comparable tasks were also observed (Kumar and Nandakumar, 2022). CLIP, functioning as both a textual and visual feature extractor, demonstrated extremely high performance on our task. We experimented with two types of fusion in case of of CLIP. Firstly, the concatenation of visual and textual features generated by CLIP was experimented with. Secondly, cross fusion for the same features was explored in which the extracted feature vectors had their outer product computed into a resulting matrix $R = p_t \otimes p_i$. Surprisingly, higher results were obtained by concatenating the features. A 3-layer classification head was implemented, utilizing RELU as its activation function.

4.4 Ensembling

Combining various models to enhance robustness, generalization, and predictive performance is a practice known as ensembling in machine learning. In our approach, hard voting is utilized, where predictions on a dataset are made by individual models within the ensemble, and the final prediction is determined through majority voting. Experiments were conducted involving the ensemble of top-k learners for each subtask, leading to the derivation of our predictions.

4.5 Experiment Settings

The training procedure was conducted using the Google Colab [4] platform for training our pipeline, which has 12.68 GB of RAM, a 14.75 GB NVIDIA Tesla T4 GPU, and Python language. Table 3 and Table 4 illustrate the hyperparameters used both in experimenting with CLIP and BERT-Based models.

5 Results

This section will expand on the result obtained through the usage of the aforementioned systems. For CLIP, HateBERT, Swin and ViT, we experimented with a variety of model sizes. Top-k Ensemble would then choose the highest k submissions to ensemble them.

[4] https://colab.google/

Hyperparameter	Value
Epochs	30
Learning Rate	2e-5
Batch Size	16
Max length	128
Optimizer	Adam
Early Stopping Patience	5
Reduce On Plateau	2
Loss Function	Dice Loss

Table 3: Training Hyperparameters for BERT-BASED.

Hyperparameter	Value
Epochs	10
Learning Rate	1e-5
Batch Size	16
Optimizer	Adam
Early Stopping Patience	5
Reduce On Plateau	2
Loss Function	Cross Entropy

Table 4: Training Hyperparameters for CLIP.

5.1 Subtask A

Table 5 illustrates the performance of the previously mentioned models on the test set. Text models demonstrated good outcomes, surpassing certain suggested multimodal models. Notably, CLIP outperforms all proposed models without requiring fine-tuning, presenting significant advantages in terms of training time. It is noteworthy that ensembling various models resulted in a marginal performance improvement, prompting inquiries about the effectiveness of an ensemble approach when compared to using only CLIP.

Model	Precision	Recall	F1-Score
RoBERTa	0.8243	0.8246	0.8245
HateBERT	0.8214	0.8186	0.8169
XLMRoBERTa	0.7676	0.7676	0.7676
Swin+HateBERT	0.7599	0.7576	0.7576
ViT+HateBERT	0.8161	0.8153	0.8157
CLIP (Cross)	0.8464	0.8448	0.8454
CLIP (Concat)	0.8546	**0.8540**	0.8543
Top-3 Ensemble	**0.8550**	0.8539	**0.8544**

Table 5: Results For Subtask A.

5.2 Subtask B

Table 6 illustrates the performance of the previously mentioned models on the test set, yet for this subtask out other multimodal approaches were not able to converge really well so unlike the first subtask they were not used for testing. Concatenating CLIP features outperformed all of its peers yet was beaten by ensembling top-3 performing models with a very small margin. This raises doubts about the effectiveness of the ensemble approach compared to utilizing only CLIP.

Model	Precision	Recall	F1-Score
RoBERTa	0.6832	0.7208	0.6960
HateBERT	0.6669	0.7479	0.6877
XLMRoBERTa	0.5866	0.5990	0.5910
CLIP (Cross)	0.7391	0.7372	0.7379
CLIP (Concat)	0.7465	0.8240	0.7671
Top-3 Ensemble	**0.7499**	**0.8273**	**0.7703**

Table 6: Results For Subtask B.

5.3 Leaderboard Results

During the evaluation phase of the shared task, we submitted our models for assessment on the test sets of both Subtask A and Subtask B. The outcomes of the tests are presented in Table 5 and Table 6, respectively. Our multimodal ensemble, which combines CLIP and BERT-based models, achieved the second place among the 7 participating teams in Subtask A. Similarly, the same model secured the second position among the 5 participating teams in Subtask B. One really intriguing direction is exploring explainable AI. In recent years, there has been a lot of approaches to explain the reasoning behind the model's predictions like Grad-Cam (Selvaraju et al., 2019), LIME(Ribeiro et al., 2016) and many others. (Chefer et al., 2021) proposed a technique for explaining transformer based models that could be adapted to our model, something that would further solidify our model's performance and open the door for many improvements as we may use such a technique for advanced error analysis.

6 Discussion & Future Work

The results obtained underscore the capability of CLIP in achieving promising outcomes for multimodal text-embedded image classification. These

findings lay a robust groundwork for future research pursuits. One avenue worth investigating involves understanding the reasons behind the limited generalization of vision models on text-embedded images. In fact, an intriguing strategy is presented in (Pramanick et al., 2021), where image attributes are extracted instead of encoded features. Another compelling approach is to delve into language models with visual understanding, such as GPT-4.

7 Conclusion

This study outlines the endeavors of our team, "AAST-NLP," in addressing the pervasive issue of using text-embedded images for hate speech and propaganda. However, it is important to note that text-embedded images also have the potential to be utilized for positive purposes. Despite their potential for misuse, as observed during the Russia-Ukraine war, the identification and mitigation of such instances are crucial, particularly in times of prolonged conflict. Our solution makes use of ensembling via hard voting based on CLIP and BERT-Based models. Our model has the potential to be used in lots of aspects as a result of its relatively high performance on both subtasks.

References

Abdul Aziz, MD. Akram Hossain, and Abu Nowshed Chy. 2023. CSECU-DSG@multimodal hate speech event detection 2023: Transformer-based multimodal hierarchical fusion model for multimodal hate speech detection. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 101–107, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17(9):e0274300.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv (Cornell University)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Cagri Toraman. 2023. Arc-nlp at multimodal hate speech event detection 2023: Multimodal methods boosted by ensemble learning, syntactical and entity features.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ursula Kristin Schmid. 2023. Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media Society*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Karanpreet Singh, Vajratiya Vajrobol, and Nitisha Aggarwal. 2023. IIC_Team@multimodal hate speech event detection 2023: Detection of hate speech and targets using xlm-roberta-base. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 136–143, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Farhan Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 151–159, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis - shared task at case 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songiln Hu. 2022. Multimodal hate speech detection via Cross-Domain knowledge transfer. *Proceedings of the 30th ACM International Conference on Multimedia*.