CSECU-DSG@ Multimodal Hate Speech Event Detection 2023: Transformer-based Multimodal Hierarchical Fusion Model For Multimodal Hate Speech Detection

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

The emergence of social media and e-commerce platforms enabled the perpetrator to spread negativity and abuse individuals or organisations worldwide rapidly. It is critical to detect hate speech in both visual and textual content so that it may be moderated or excluded from online platforms to keep it sound and safe for users. However, multimodal hate speech detection is a complex and challenging task as people sarcastically present hate speech and different modalities i.e., image and text are involved in their content. This paper describes our participation in the CASE 2023 multimodal hate speech event detection task. In this task, the objective is to automatically detect hate speech and its target from the given text-embedded image. We proposed a transformer-based multimodal hierarchical fusion model to detect hate speech present in the visual content. We jointly fine-tune a language and a vision pre-trained transformer models to extract the visual-contextualized features representation of the text-embedded image. We concatenate these features and fed them to the multi-sample dropout strategy. Moreover, the contextual feature vector is fed into the BiLSTM module and the output of the BiLSTM module also passes into the multi-sample dropout. We employed arithmetic mean fusion to fuse all sample dropout outputs that predict the final label of our proposed method. Experimental results demonstrate that our proposed method obtains competitive performance and ranked 5th among the participants.

1 Introduction

Nowadays social media increasingly become popular means of information sharing because people consistently present their concepts, opinions, thoughts, and breaking news using various platforms including Twitter, Facebook, Reddit, and Instagram as their real-time behaviour and practical features. Online abuse and the spreading of negativity are common practices and important societal problem that is highly correlated with the emergence of social media platforms (Parihar et al., 2021). Analyzing and extracting social media information have various benefits as it promotes a safer online platform, reduces online harassment and cyberbullying, and reduces harmful and false information. However, detecting hate space on social media content is a complex and challenging task as people express their information sarcastically i.e., memes, the multifaceted nature of content, and multiple modalities are involved. Researchers consider hate speech detection as the text-only task at the commencement stage (Djuric et al., 2015; Badjatiya et al., 2017; Watanabe et al., 2018). However, this practice is not effective as people share text-embedded pictures or memes as well which helps to understand the real scenario of the content. Essentially, hate speech detection slowly moves to the visual-textual format named multimodal hate speech detection (Sabat et al., 2019; Thapa et al., 2022; Chhabra and Vishwakarma, 2023). Multimodal hate speech is now one of the most popular tasks and developed various methods (Cai et al., 2019; Gomez et al., 2020; Zhu et al., 2022). Facebook AI introduce hateful meme challenges (Kiela et al., 2020) and various teams proposed state-of-the-art methods (Velioglu and Rose, 2020; Lippe et al., 2020). Velioglu and Rose (2020) proposed a winning approach where they utilize VisualBERT and ensemble learning to detect hateful memes. Gomez et al. (2020) introduced a large-scale multimodal hate speech dataset of multimodal publication from Twitter and provided various unimodal and multimodal baseline methods. Yang et al. (2022) proposed a cross-domain knowledge transfer (CDKT) framework for the multimodal hate speech detection task where they used a vision-language transformer as the backbone of the proposed approach. Recently, the Russia-Ukraine issue has been a significant topic of discussion on

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 101–107, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_014

social media platforms and people present their opinions and thoughts on social media. Bhandari et al. (2023) proposed a multimodal hate speech detection dataset, CrisisHateMM based on the Russia-Ukraine crisis on social media. They provide a multimodal analysis of directed and undirected hate speech in text-embedded pictures from the Russia-Ukraine conflict. Thapa et al. (2023) introduce a shared task at CASE 2023 based on the CrisisHateMM dataset where the participant's system needs to detect hate speech and target from the given text-embedded image in a multimodal setting. To tackle this task we propose a transformer-based multimodal hierarchical fusion approach with the BiLSTM module and the multi-sample dropout strategy. Our system obtained competitive performance and ranked 5th in both sub-tasks.

We organize the rest of the paper as follows: In **Section 2**, we provide detailed descriptions of the task and dataset. **Section 3** describes our proposed system in the CASE 2023 task 4: multimodal hate speech event detection task to automatically detect hate speech and target. In **Section 4**, we present our proposed system design with parameter settings and conduct the results and component analysis. Finally, we conclude with some future directions in **Section 5**.

2 Task and Dataset Description

2.1 Task Description

The task aims to detect hate speech in text-embedded images on social media and the internet based on the topic of the Russia-Ukraine war. Text-embedded images were extensively used, both by the Russian and Ukrainian sides, to disseminate propaganda and hate speech during the Russia-Ukraine war. In this task, organizers featured two subtasks focusing on detecting hate speech and its target. In subtask A, the objective is to detect whether a given text-embedded image is hateful or not. Subtask B aims to detect the targets of hate speech in a given hateful text-embedded image.

2.2 Dataset Description

The organizers used a benchmark dataset CrisisHateMM (Bhandari et al., 2023) to evaluate the performance of the participants' systems at the CASE 2023 shared task 4 [1] (Thapa et al., 2023) to detect hate speech in text-embedded pictures. The dataset is collected from social media platforms including Twitter, Reddit, and Facebook based on the Russia-Ukraine conflict. The dataset comprises 4486 and 2428 text-embedded images for subtask A and subtask B, respectively. Subtask A comprised 3600 train, 443 dev, and 443 test text-embedded images and Subtask B consisted of 1942 train, 244 dev, and 242 test text-embedded images. The dataset statistics of subtask A: hate speech event detection and subtask B: target detection are presented in Table 1 based on each task's labels. For subtask B, text-embedded images are annotated for community, individual and organization targets whereas subtask A is annotated for the hate and non-hate labels. Moreover, texts are extracted from the text-embedded images using OCR with the Google Vision API [2].

3 Proposed Framework

Transformers models learn the necessary information about the relationship between words effectively. We employed the pre-trained transformers model with the BiLSTM module and a training strategy to detect the hate speech of text-embedded images in a multimodal setting. The overview of our proposed transformer-based framework is delineated in Figure 1.

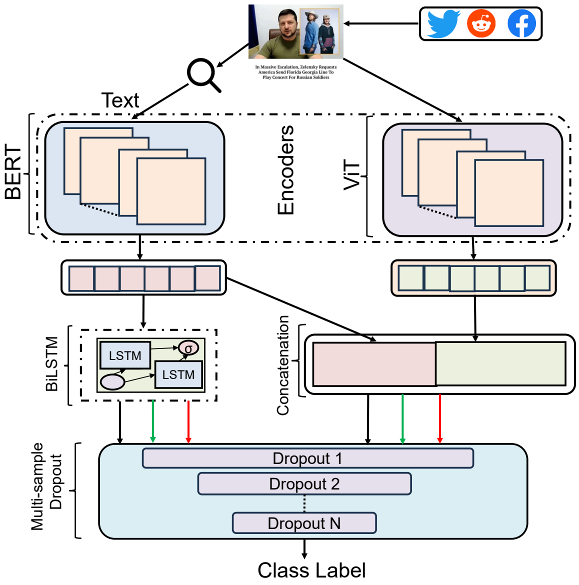
In Massive Escalation, Zelensky Requests America Send Florida Georgia Line To Play Concert For Russian Soldiers
Text
BERT
Encoders
ViT
BiLSTM
LSTM
LSTM
Concatenation
Multi-sample Dropout
Dropout 1
Dropout 2
Dropout N
Class Label

Figure 1: Overview diagram of our proposed method for multimodal hate speech detection

Given a text-embedded image, we extract the text from the image. We fed the extracted text and

[1] https://github.com/therealthapa/case2023_task4

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

Category	Subtask A		Subtask B		
	Hate	Non-Hate	Individual	Community	Organization
Train	1,942	1,658	823	335	784
Dev	243	200	102	40	102
Test	243	200	102	42	98
Total	2,428	2,058	1,027	417	984

Table 1: The statistics of the used dataset in CASE 2023 shared task 4 across all subtasks.

text-embedded image into a language model and a vision pre-train transformer model to extract the visual-contextualized embedding features, respectively. We concatenate these feature vectors to get the multimodal unified representation of the image-text pair. Although, contextualized embedding features are fed into the BiLSTM module to learn the long-term contextual dependency that helps the model to effectively capture the hate information present in the context. A multi-sample dropout strategy is employed on top of both multimodal and BiLSTM module outputs to improve the generalization ability and robustness leading to performance enhancement. Later, we utilise an arithmetic mean fusion to get the final prediction of our proposed approach.

3.1 Transformers Model

We fine-tuned the BERT transformers model to extract the contextualized features representation of text. ViT transformers model is employed to capture the visual information in the given image.

3.1.1 BERT

BERT (Devlin et al., 2018) stands for bidirectional encoder representations from transformers, is a new method of pre-training sentence representations which achieves state-of-the-art results on many NLP tasks including question-answering, text classification, and sentence-pair regression. It is trained on a large corpus of unlabelled text which includes the entire Wikipedia (that's about 2500 million words) and a book corpus (800 million words). We take advantage of the BERT fast tokenizer and *bert-base-uncased* model with fine-tuning to learn a 768-dimensional textual feature vector of the extracted text from the text-embedded image. It is composed of 12 transformer blocks, a hidden size of 768, and 110M parameters with a vocabulary of 30K tokens in the embedding layer.

3.1.2 ViT

The vision transformer (ViT) (Dosovitskiy et al., 2020) is a transformer encoder model (BERT-like) pre-trained on a large collection of images in a self-supervised fashion. ViT split an image into patches and flatten the patches to produce lower-dimensional linear embeddings from the flattened patches. Add positional embeddings and feed the sequence as an input to a standard transformer encoder. Image patches are the sequence tokens like words. The encoder block is identical to the original transformer architecture. It is utilized ImageNet-1k, at a resolution of 224x224 pixels and fixed-size patches with a resolution of 16x16. We employ the ViT model's *facebook/dino-vitb16* checkpoint trained using the DINO method to extract the visual features of the given image.

3.2 BiLSTM Module

We employed a BiLSTM layer (Brueckner and Schulter, 2014) on top of the BERT model's textual representation that helps the model capture and enriches textual information presented in the extracted text. The BiLSTM module is strong enough in capturing long-range dependencies in sequential data that result in more informative feature representations. Multimodal hate speech detection is a text-dominant task hence BERT transformer model with BiLSTM-based effective textual representation could benefit in understanding hate information present in the text-embedded image that will lead to the improved performance of unified multimodal architecture. Here, BiLSTM can effectively learn the long-term contextual dependency from the BERT transformer model's textual representation in our approach.

3.3 Multi-sample Dropout Strategy

Different training strategies improved the performance of the transformers model. In this paper, we use a multi-sample dropout training strategy (Inoue, 2019). To improve the accuracy of the transformer-based multimodal hierarchical fusion network, we utilise the multi-sample dropout technique. Although, it improves the generalization ability and accelerates the training of base model (Inoue, 2019). We employed this technique in two stages. Firstly, we employ multi-sample dropout after the multimodal features vectors. Secondly, we fed the BiLSTM module output to the multi-sample dropout. This hierarchical fusion helps the model to learn the context effectively. In multi-sample dropouts, we duplicate the features vector of the multimodal and BiLSTM module output after the dropout layer, while sharing the weights among these duplicated fully connected layers. To obtain the final loss, we calculate the loss for each sample, and then the sample losses are leveraged using the arithmetic mean-based fusion.

Parameter	Optimal Value
Learning rate	3e-5
Max-len	128
Number of epochs	5
Batch size	8
Manual seed	42
Dropout	0.6, 0.7,0.8

Table 2: Proposed model hyperparameters settings for CASE-2023 task 4 shared task.

4 Experimentals and Evaluations

4.1 Experimental Settings

We now describe the details of our experimental settings and the hyper-parameter settings with the fine-tuning strategy that we have employed to design our proposed multimodal approach for the CASE 2023 shared task 4. We finetune state-of-the-art Huggingface (Wolf et al., 2019) transformer models including BERT [3] and DINO Vit [4] model for this task. We used all models as the base size in this work. We concatenate the training and development data during the model training phase. We implement our proposed method using PyTorch (Paszke et al., 2019). We used the CUDA-enabled GPU of the Google Colaboratory (Bisong and Bisong, 2019) platform and set the manual seed = 42 to generate reproducible results. We obtained the optimal parameter settings of our proposed model based on the performance of the development set which is articulated in Table 2. We use a multi-sample dropout training strategy on top of the unified representation of multimodal and multigenre tasks. To determine the optimal dropout values, we searched over the set $\{0.1, 0.2, \cdots, 0.9\}$ and found the best dropout range was 0.6 to 0.8 based on our experimental results on the development set. We used the default settings for the other parameters.

4.2 Evaluation Measures

To evaluate the performance of participants' lexical complexity prediction systems, CASE 2023 task 4 organizers used different strategies and metrics for sub-task A and sub-task B (Thapa et al., 2023). For both sub-task, standard evaluation metrics including precision, recall, F1-score and accuracy were applied to estimate the performance of a system. However, the macro-averaged F1 score is considered as the primary evaluation measure for both subtasks of this task.

4.3 Results and Analysis

In this section, we analyze the performance of our proposed CSECU-DSG system in the CASE 2023 multimodal hate speech event detection shared task. We used the full training set and validation set for training our proposed model and also the validation set for hyperparameter tuning.

The comparative performance of our proposed CSECU-DSG system on subtask A hate speech detection test data against other selected participants' systems is presented in Table 3. We have seen that our proposed method achieved a 0.8248 F1 score and 0.8262 accuracy and ranked 5th in sub-task A based on the macro-averaged F1 score. Our proposed approach surpasses the CLIP model baseline (Bhandari et al., 2023) method by 8.23% and obtains competitive performance. This validates the effectiveness of our proposed method in the multimodal hate speech detection task.

The comparative performance of our proposed CSECU-DSG system on subtask B target detection against other selected participants' systems and baseline method is presented in Table 4. In the target detection task, our method achieved a 0.6530 F1 score and a 0.6901 accuracy score. Our proposed method outperforms the baseline method by 5.82%

[3] https://huggingface.co/bert-base-uncased

[4] https://huggingface.co/facebook/dino-vitb16

Team	Recall	Precision	F1 score	Accuracy	Rank
CSECU-DSG	0.8252	0.8244	0.8248	0.8262	5th
Participants system performance on subtask A					
ARC-NLP (Sahin et al., 2023)	0.8567	0.8563	0.8565	0.8578	1st
bayesiano98 (Thapa et al., 2023)	0.8562	0.8528	0.8528	0.8233	2nd
DeepBlueAI (Thapa et al., 2023)	0.8356	0.8335	0.8342	0.8352	4th
Avanthika (Thapa et al., 2023)	0.7878	0.7881	0.7880	0.7901	7th
rabindra.nath (Thapa et al., 2023)	0.7768	0.7842	0.7788	0.7833	9th
GT (Thapa et al., 2023)	0.5219	0.5219	0.5219	0.5260	11th
Baseline (CLIP) (Bhandari et al., 2023)	-	-	0.7860	0.7980	-

Table 3: Comparative results with other selected participants and baseline on Subtask A: Hate speech detection. The teams are ranked based on the macro-averaged F1 score. Our team name is CSECU-DSG.

Team	Recall	Precision	F1 score	Accuracy	Rank
CSECU-DSG	0.6525	0.6575	0.6530	0.6901	5th
Participants system performance on subtask B					
ARC-NLP (Sahin et al., 2023)	0.7636	0.7637	0.7634	0.7934	1st
bayesiano98 (Thapa et al., 2023)	0.7330	0.7554	0.7410	0.7727	2nd
IIC_Team (Thapa et al., 2023)	0.6894	0.7105	0.6973	0.7231	3rd
DeepBlueAI (Thapa et al., 2023)	0.6462	0.6648	0.6525	0.6983	6th
Ometeotl (Thapa et al., 2023)	0.5648	0.6793	0.5677	0.6405	7th
ML_Ensemblers (Thapa et al., 2023)	0.4444	0.4888	0.4332	0.5289	9th
Baseline (CLIP) (Bhandari et al., 2023)	-	-	0.6150	0.6840	-

Table 4: Comparative results with other selected participants and baselines on Subtask B: Target detection. The teams are ranked based on the macro-averaged F1 score. Our team name is CSECU-DSG.

and is ranked 5th in this task leaderboard [5] in terms of primary evaluation measure macro-averaged F1 score. This validates the potency and applicability of our proposed method in the target detection task.

4.4 Discussion

To estimate the contribution of the BiLSTM module and multi-sample dropout training strategy in our proposed approach for multimodal hate speech event detection task, we performed the component ablation study. In this regard, we first removed the multi-sample dropout training strategy, the BiLSTM module, and both multi-sample dropout strategies at each time and repeated the experiment. The results of our ablation study are reported in Table 5. We first report our team's CSECU-DSG performance and then the other method's performance based on the macro-averaged F1 score. It shows that when removing the multi-sample dropout strategy the results decrease on average 1% and removing the BiLSTM module from the proposed method leads to a decrease in the results of 1.3% in terms

Method	Subtask A	Subtask B
CSECU-DSG	0.8248	0.6530
- MSD	0.8164	0.6462
- BiLSTM	0.8143	0.6441
- MSD+BiLSTM	0.8065	0.6207

Table 5: The ablation study of our proposed method based on the test dataset in CASE 2023 shared task 4 across all subtasks. The result is reported in terms of primary evaluation measure macro-averaged f1 score. MSD stand for multi-sample dropout.

[5] https://codalab.lisn.upsaclay.fr/competitions/13087#results

of macro-averaged F1 score. We observed 2.2% performance decreases in subtask A and 3.7% performance decreases in subtask B based on macro-averaged F1 score when we remove both the BiLSTM module and multi-sample dropout strategy at a time which deduced the contribution of the multi-sample dropout training strategy and BiLSTM module components in our model.

5 Conclusion and Future Work

In this paper, we present an approach to automatically identify hate speech in multimodal settings using fine-tuned transformers models fusion architecture. We employ a BiLSTM module on top of the language model to handle the long-term dependencies present in the context. Moreover, we use the multi-sample dropout training strategy to speed up training and get better generalization ability. Experimental results demonstrated the efficacy of our proposed transformer-based method, where the hierarchical fusion of transformer variants with the BiLSTM module and multi-sample dropout prediction helped us to obtain competitive performance and ranked 5th in both subtasks in the CASE 2023 shared task 4: multimodal hate speech event detection.

Further research will be conducted on other large transformers models with a unified architecture of two or more. However, the classes of the dataset are imbalanced, so the weighted average fusion strategy of different models may be exploiting better context for hate speech from multimodal content effectively.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

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*.

Ekaba Bisong and Ekaba Bisong. 2019. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64.

Raymond Brueckner and Björn Schulter. 2014. Social Signal Classification Using Deep BLSTM Recurrent Neural Networks. In *2014 IEEE International Conference on Coustics, Speech and Signal Processing (ICASSP)*, pages 4823–4827. IEEE.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

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. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

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. *arXiv preprint arXiv:2307.13829*.

Surendrabikram Thapa, Farhan Ahmad 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 (CASE)*.

Surendrabikram Thapa, Aditya Shah, Farhan Ahmad Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

Jiawen Zhu, Roy Ka-Wei Lee, and Wen Haw Chong. 2022. Multimodal zero-shot hateful meme detection. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 382–389.