CLTL@Multimodal Hate Speech Event Detection 2024: The Winning Approach to Detecting Multimodal Hate Speech and Its Targets

Yeshan Wang
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
yestin-wang@outlook.com

Ilia Markov
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
i.markov@vu.nl

Abstract

In the context of the proliferation of multimodal hate speech related to the Russia-Ukraine conflict, we introduce a unified multimodal fusion system for detecting hate speech and its targets in text-embedded images. Our approach leverages the Twitter-based RoBERTa and Swin Transformer V2 models to encode textual and visual modalities, and employs the Multilayer Perceptron (MLP) fusion mechanism for classification. Our system achieved macro F1 scores of 87.27% for hate speech detection and 80.05% for hate speech target detection in the Multimodal Hate Speech Event Detection Challenge 2024, securing the 1st rank in both subtasks. We open-source the trained models at https://huggingface.co/Yestin-Wang

1 Introduction

In the ever-evolving digital age, social media platforms have emerged as pivotal arenas for information exchange and social interaction. This surge in online engagement, while fostering connectivity and the exchange of ideas, has also led to a rise in online abuse, including the spread of hate speech. Hate speech, commonly defined as communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000), has emerged as a significant societal issue. The complexity of identifying hate speech is further amplified by the multimodal nature of online content, often in the form of text-embedded images. These images, which combine visual and textual elements, are a prevalent mode of expression on social media platforms (Shang et al., 2021). The challenge of detecting hate speech in text-embedded images arises from the multimodal nature of the content, where textual cues are intertwined with visual content. Traditional unimodal models, which focus solely on text or image classification, fall short in effectively interpreting the nuanced and often context-dependent nature of hate speech in these multimodal scenarios (Kiela et al., 2020). Therefore, there is a critical need for advanced multimodal models that can effectively integrate and analyze both textual and visual information to accurately identify hate speech in text-embedded images.

In light of this need, the Multimodal Hate Speech Event Detection Challenge[1] at CASE 2024 (Thapa et al., 2024) provides a platform for developing and evaluating models capable of detecting hate speech in text-embedded images, concerning politically controversial topics related to the Russia-Ukraine War. This task builds on the 2023 iteration of this shared task (Thapa et al., 2023), which includes subtasks aimed at not only determining the presence of hate speech in such images but also identifying the targets of hateful content, whether they are individuals, organizations, or communities. Most of the participating teams from previous year had employed supervised approaches based on unimodal transformer models (e.g., BERT, XLM-Roberta, etc.) (Armenta-Segura et al., 2023; Singh et al., 2023) or methods based on feature engineering (e.g., lexical features, named entities, amongst others) and ensemble learning strategies (Sahin et al., 2023). However, these approaches often required complex feature engineering and specialized model structure for specific subtasks, which makes it challenging to generalize across different subtasks, such as detecting hate speech and its targets (Thapa et al., 2023).

We introduce an unified multimodal architecture for both hate speech and target detection tasks. Our approach employs Twitter-based RoBERTa (Loureiro et al., 2023) and Swin Transformer V2 models (Liu et al., 2022) to extract features for

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

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 73–78

encoding textual and visual content and concatenates them via the Multilayer Perceptron (MLP) fusion technique. Without the need for feature engineering, our system achieved first place in both subtasks, outperforming the previous year's top team by 1.62% F1 score in subtask A and 3.71% F1 score in subtask B, respectively.

2 Related Work

In the intersection of natural language processing and computer vision, the detection of hate speech in multimodal content, especially in text-embedded images, has increasingly attracted scholarly attention. This trend is driven by both the development of innovative multimodal methodologies and the creation of extensive datasets (Bhandari et al., 2023; Fersini et al., 2022; Pramanick et al., 2021; Suryawanshi et al., 2020). A pivotal advancement was the Hateful Memes Challenge at NeurIPS 2020 (Kiela et al., 2020), which provided an open-source dataset comprising 10,000 meme examples, fostering a competitive environment for developing state-of-the-art methods. The winning approach by Zhu (2020) combined VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020), VILLA (Gan et al., 2020) and ERNIE-ViL (Yu et al., 2021) through ensemble learning, demonstrating enhanced capability in multimodal hate speech detection.

Research on multimodal hate speech detection has predominantly focused on multimodal fusion approaches, essential for handling the dual-modality of text-embedded images. These methodologies are typically divided into early fusion, which combines text and visual features at the initial stages using deep fusion encoders with cross-modal attention (Atrey et al., 2010), and late fusion, which employs separate processing of image and text modalities before merging them at a decisive alignment stage (Li et al., 2022). Recent studies employed novel feature extraction techniques to improve classification efficacy. For instance, Zhou et al. (2021) proposed an image captioning-based feature extraction method, generating descriptive texts from multimodal memes. Blaier et al. (2021) showed that incorporating caption features during model fine-tuning improves the performance of various multimodal models for hateful meme detection.

The scope of multimodal hate speech detection is continually widening to cover a wide range of hate speech triggering events, such as presidential elections (Suryawanshi et al., 2020), the COVID-19 pandemic (Pramanick et al., 2021), and geopolitical conflicts like the Russia-Ukraine War (Thapa et al., 2022). Initiatives to label and detect harmful text-embedded images in these specific contexts contribute to a deeper understanding of how multimodal hate speech manifests itself during various significant events.

3 Dataset & Task Description

3.1 Dataset Description

The dataset used for the shared task is CrisisHateMM (Bhandari et al., 2023). It comprises 4,723 text-embedded images, reflecting diverse social media discourses related to the Russia-Ukraine conflict. The dataset is meticulously compiled from popular social media platforms such as Twitter, Reddit, and Facebook. Each item in the dataset comprises an original image file alongside its extracted textual content, obtained via OCR technology using the Google Vision API[2].

Subtask	Classes	Train	Eval	Test
Subtask A	Hate	1,942	243	243
	No Hate	1,658	200	200
Subtask B	Individual	823	102	102
	Community	335	40	42
	Organization	784	102	98

Table 1: Dataset statistics: number of instances across different subtasks.

For both subtasks, the dataset is split into training, evaluation, and test sets. Table 1 provides the number of instances in each set. Notably, the test set labels remain undisclosed during the challenge phase to ensure an unbiased performance evaluation.

3.2 Subtask A: Hate Speech Detection

Subtask A focuses on identifying whether a given text-embedded image contains hate speech or not, corresponding to the binary classification problem. There are 2,428 text-embedded images labeled as containing hate speech, and 2,058 non-hate speech examples in the dataset, which is divided into 3,600 training, 443 evaluation, and 443 testing instances. This division ensures that there is the same number of instances per class in the evaluation and test sets.

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

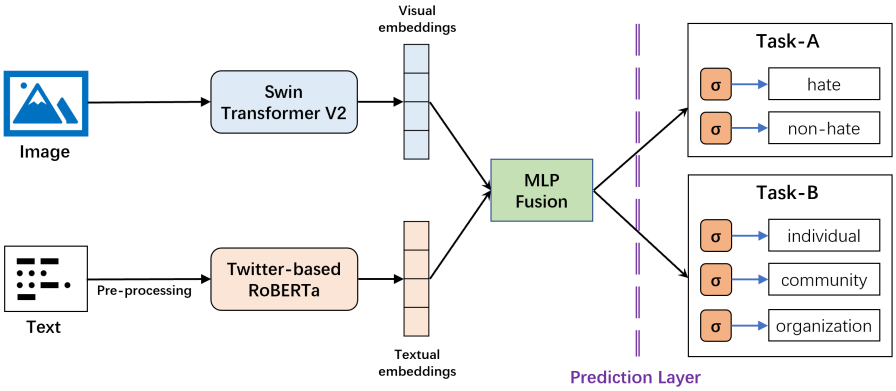

Figure 1: An overview of the CLTL's system.

3.3 Subtask B: Target Detection

Subtask B aims to identify the targets of hate speech within 2,428 text-embedded hateful images. Each text-embedded image in this subtask is annotated for targeting specific groups or entities: "community", "individual" or "organization", which is viewed as a multi-class, single-label classification problem.

4 Methodology

Our approach is based on a multimodal architecture that integrates large-scale pre-trained models, Twitter-based RoBERTa (Loureiro et al., 2023) and Swin Transformer V2 (Liu et al., 2022), to extract contextualized embeddings from textual and visual inputs, respectively. These embeddings are then concatenated using the Multilayer Perceptron (MLP) fusion module (Shi et al., 2021) for classifying each instance into one of the predefined categories. Our model is universally applicable to both subtasks A and B, differing only in the output layer, as depicted in Figure 1.

4.1 Text Preprocessing

The provided dataset comes with the textual data extracted from text-embedded images via Google OCR Vision API. We applied simple preprocessing steps, which involve removing URLs, username mentions (i.e., @username), and emojis using regular expressions, followed by setting the text truncation length to 512 tokens. These preprocessing steps are commonly applied when dealing with social media data (Gupta and Joshi, 2017).

4.2 Transformers Models

4.2.1 Twitter-based RoBERTa

The Twitter-based RoBERTa model (Loureiro et al., 2023) is a RoBERTa-large model (Liu et al., 2019) trained on a large corpus of 154M tweets covering the periods between 2018-01 and 2022-12, possibly covering tweets related to the Russia-Ukraine conflict as well. Considering that the properties of tweets are to some extent similar to the properties of texts embedded in images in our data: both are short texts, containing informal language, abbreviations and slang specific to social media, a domain-specific large language model is expected to be more suitable for encoding textual input. The Twitter-based RoBERTa model is publicly available via the Hugging Face Transformer API[3].

4.2.2 Swin Transformer V2

Swin Transformer V2 (Liu et al., 2022) is an improved version of the Swin Transformer (Liu et al., 2021), which employs a window-based attention mechanism for efficient image processing across various scales and resolutions by partitioning the image into non-overlapping patches and processing these sequentially at each stage. This approach mitigates the computational and memory burden issues of large-scale image processing in traditional transformer architectures that apply global self-attention mechanisms across the entire image. In our experiments, we use the TIMM framework implementation of the Swin Transformer V2 model[4]. The model was pretrained on the ImageNet-1k dataset, containing a collection of 1.2 million labeled images with one thousand object categories (Russakovsky et al., 2015).

4.3 MLP Fusion & Prediction

Our fusion strategy entails the concatenation of text and image embeddings through the Multilayer

[3] https://huggingface.co/cardiffnlp/twitter-roberta-large-2022-154m

[4] https://huggingface.co/timm/swinv2_base_window8_256.ms_in1k

Team	Recall	Precision	F1-score	Rank
Ours (CLTL)	87.37	87.20	87.27	1st (2024)
ARC-NLP	85.67	85.63	85.65	1st (2023)
AAST-NLP	85.46	85.40	85.43	2nd (2024)
bayesiano98	85.61	85.28	85.28	2nd (2023)
Baseline (CLIP)	-	-	78.60	-

Table 2: Performance comparison for the baseline approach (Bhandari et al., 2023) and top-performing teams in 2023/24 multimodal hate speech detection task (Thapa et al., 2023, 2024).

Perceptron (MLP) fusion module (Shi et al., 2021), where the top vector representations from different models are pooled (concatenated) into a single vector. A prediction layer is added at the end to perform classification: for subtask A, the sigmoid outputs a single value to yield a probability of hate speech presence. For subtask B, each target category (individual, community, and organization) has a separate sigmoid function that outputs the corresponding probability.

5 Experimentals and Evaluation Results

5.1 Experimental Settings

Our multimodal classification system was developed using the PyTorch framework and AutoGluon library (Shi et al., 2021) for a robust and flexible implementation. We fine-tuned Twitter-based RoBERTa and Swin Transformer V2 models on the training data with the following hyperparameters: a base learning rate of 1e-4, decay rate of 0.9 using cosine decay scheduling, batch size of 8, and a manual seed of 0 for reproducibility. The models were optimized using the AdamW optimizer for up to 10 epochs, or until an early stopping criterion was met to prevent overfitting. After fine-tuning, the models were assessed on the evaluation set. All experiments were conducted on the Google Colaboratory platform with a NVIDIA A100 GPU, taking approximately 25 minutes for subtask A and 20 minutes for subtask B.

5.2 Results & Discussion

The official evaluation metric to score participating systems was macro-averaged F1 score as the test set is imbalanced. Table 2 and 3 showcase the comparative performance of the CLTL team's system in addressing the challenging tasks of multimodal hate speech detection (subtask A) and target detection (subtask B), reflect the superior performance of our system in comparison to the baseline approach (Bhandari et al., 2023) and other top-ranked participating systems (Thapa et al., 2023, 2024).

In subtask A, our system obtained an F1 score of 87.27%, achieving the top rank on the leaderboard. This performance represents a substantial improvement over the previous year's winning entry (Sahin et al., 2023), with an increase of 1.62% in F1 score. Notably, our system excelled across all classification metrics within the test results, and outperformed the CLIP model baseline approach by 8.67% in terms of F1 score, highlighting the robustness of our approach for multimodal hate speech detection.

In subtask B, our system again led the rankings, achieving an F1 score of 80.05%. This is a notable advancement of 18.55% over the F1 score of the baseline approach, which validates the effectiveness of our system in identifying the specific targets of multimodal hate speech.

The success of our system across both subtasks could be potentially attributed to several factors. The extensive pre-training on large volumes of textual and visual content has been instrumental, partially due to the Twitter-based RoBERTa's domain-specific knowledge of social media discourse, and the Swin Transformer V2's proficiency in visual understanding. The subsequent fine-tuning on the CrisisHateMM training set has further enhanced the system's capacity for classifying multimodal hateful content. Moreover, the concatenation of text and image modalities via the MLP fusion module, has proven effective in capturing the complex interplay between textual and visual cues inherent in multimodal hate speech and its targets.

6 Conclusion

In this work, we introduced the multimodal architecture designed by CLTL team for the Multimodal Hate Speech Event Detection Challenge 2024. Leveraging the Twitter-based RoBERTa and Swin Transformer V2 for feature extraction and

Team	Recall	Precision	F1-score	Rank
Ours (CLTL)	79.07	81.48	80.05	1st (2024)
AAST-NLP	74.65	82.40	76.71	2nd (2024)
ARC-NLP	76.36	76.37	76.34	1st (2023)
bayesiano98	75.54	73.30	74.10	2nd (2023)
Baseline (CLIP)	-	-	61.50	-

Table 3: Performance comparison for the baseline approach (Bhandari et al., 2023) and top-performing teams in 2023/24 multimodal hate speech target detection task (Thapa et al., 2023, 2024).

employing the MLP fusion mechanism, our system achieved the top rank with the highest macro F1 score on both subtasks, which sets a new state-of-the-art in detecting hate speech and its targets on the CrisisHateMM dataset. In future work, we aim to refine our approach by experimenting with advanced fine-tuning strategies such as parameter-efficient fine-tuning (PEFT), using larger multimodal datasets to improve the generalization capabilities of our approach across diverse social media domains.

References

Jesus Armenta-Segura, César Jesús Núñez-Prado, Grigori Olegovich Sidorov, Alexander Gelbukh, and Rodrigo Francisco Román-Godínez. 2023. Ometeotl@multimodal hate speech event detection 2023: Hate speech and text-image correlation detection in real life memes using pre-trained BERT models over text. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 53–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Syst.*, 16(6):345–379.

Aashish Bhandari, Siddhant B. 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1994–2003.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9350–9358, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 6616–6628. Curran Associates, Inc.

Itisha Gupta and Nisheeth Joshi. 2017. Tweet normalization: A knowledge based approach. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pages 157–162.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965.

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. *arXiv/1907.11692*.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin

Transformer V2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and José Camacho-Collados. 2023. Tweet insights: A visualization platform to extract temporal insights from twitter. *arXiv/2308.02142*.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

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. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 71–78, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. AOMD: An analogy-aware approach to offensive meme detection on social media. *Information Processing Management*, 58(5):102664.

Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. 2021. Multimodal AutoML on structured tables with text fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*.

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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

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, Hari 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.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3208–3216.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv/2012.08290*.