ARC-NLP at Multimodal Hate Speech Event Detection 2023: Multimodal Methods Boosted by Ensemble Learning, Syntactical and Entity Features

Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, Cagri Toraman
Aselsan Research Center, Ankara, Turkiye
{ucsahin, ekucukkaya, ogozcelik, ctoraman}@aselsan.com.tr

Abstract

Text-embedded images can serve as a means of spreading hate speech, propaganda, and extremist beliefs. Throughout the Russia-Ukraine war, both opposing factions heavily relied on text-embedded images as a vehicle for spreading propaganda and hate speech. Ensuring the effective detection of hate speech and propaganda is of utmost importance to mitigate the negative effect of hate speech dissemination. In this paper, we outline our methodologies for two subtasks of Multimodal Hate Speech Event Detection 2023. For the first subtask, hate speech detection, we utilize multimodal deep learning models boosted by ensemble learning and syntactical text attributes. For the second subtask, target detection, we employ multimodal deep learning models boosted by named entity features. Through experimentation, we demonstrate the superior performance of our models compared to all textual, visual, and text-visual baselines employed in multimodal hate speech detection. Furthermore, our models achieve the first place in both subtasks on the final leaderboard of the shared task.

1 Introduction

The Russia-Ukraine War has been a long and bitter conflict that has caused a lot of division and tension among people. Unfortunately, hate speech has played a big role in this war, spreading negativity, fueling hatred, and making the situation even more volatile. It is important to find ways to detect and combat hate speech in order to promote unity and peace.

Deep learning models are increasingly being employed in multimodal hate speech detection (Parihar et al., 2021; Thapa et al., 2022; Boishakhi et al., 2021; Gomez et al., 2020; Yang et al., 2019; Perifanos and Goutsos, 2021; Rana and Jha, 2022; Vijayaraghavan et al., 2021; Sabat et al., 2019; Madukwe et al., 2020; Kiela et al., 2020). These models leverage the power of neural networks to process and analyze complex data consisting of text, images, and videos, allowing them to capture the nuances and context of online content. By combining various modalities, such as textual and visual contents, these models can better understand the overall meaning and intent behind the shared information. They learn from large amounts of labeled data, enabling them to identify patterns and distinguish between genuine information and harmful content, including hate speech and misinformation (Toraman et al., 2022a). With their ability to integrate multiple modalities, deep learning models are playing a vital role in combating online abuse, fostering safer digital environments, and promoting responsible information dissemination.

This study addresses the challenge of combating hate speech using multiple modalities, specifically focusing on the shared task of Multimodal Hate Speech Event Detection at CASE 2023 (Thapa et al., 2023). In the shared task, Subtask A requires determining whether a text-embedded image contains hate speech. To address this, we propose a novel ensemble model that merges predictions from a multimodal deep learning model and multiple text-based tabular models which are trained with various syntactical features. On the other hand, for Subtask B, the goal is to identify the target of hate speech in a text-embedded image and classify it into the categories of “Individual”, “Community”, or “Organization”. To tackle this challenge, we introduce a novel multimodal deep learning model. We train a multimodal deep learning model and then combine its embeddings with named entity features, which are then used as input to train a new fusion model. Through experimentation, we show that our proposed models achieve superior classification performance compared to the multimodal hate speech detection baselines. Notably, our proposed models achieve the highest rank on

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 71–78, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_010

Subtask	Problem	Labels	#Text-embedded Images		
			Train	Eval	Test
A	Hate Speech	Hate	1,942	243	443
		Non-Hate	1,658	200	
B	Target	Individual	823	102	242
		Community	335	40	
		Organization	784	102	

Table 1: Dataset for the shared task on Multimodal Hate Speech Event Detection at CASE 2023. Numbers of text-embedded images in the train, evaluation and test sets for both Subtask A and B are given. Labels of the test set examples are not shared.

the final leaderboard for both subtasks in the shared task.

2 Dataset & Task

The shared task on Multimodal Hate Speech Event Detection at CASE 2023[1] consists of two distinct subtasks: Subtask A and B. The details of each subtask are presented in Table 1 along with the number of text-embedded images in the training, evaluation and test sets. It is important to note that the labels of the test set examples are not disclosed to the participants during the shared task. These labels are reserved for calculating the final prediction performance, which determines the leaderboard rankings upon completion of the shared task. Furthermore, text within the images are extracted using OCR with Google Vision API[2].

2.1 Subtask A: Hate Speech Detection

In Subtask A, it is aimed to determine the presence or absence of hate speech within text-embedded images (Thapa et al., 2022). The dataset specifically designed for this subtask includes annotated examples that indicate the existence of hate speech (Bhandari et al., 2023). The dataset features two distinct labels: “Hate Speech” and “No Hate Speech”.

2.2 Subtask B: Target Detection

Subtask B aims to identify the targets of hate speech within a given hateful text-embedded image (Thapa et al., 2022). The dataset provided for this subtask includes labels categorizing the hate speech targets into “Individual”, “Community”, and “Organization” (Bhandari et al., 2023).

Feature	Count
Word counts	1
Character counts	1
Capital ratio	1
Digit ratio	1
Special character ratio	1
White space ratio	1
Symbol (!, ?, @, %, *, $, &, #, ., :, /, -, =) ratios	13
Symbol counts	13
Lowercase ratio	1

Table 2: Syntactical features used in our proposed model for Subtask A.

3 Methodology

In this section, we describe our proposed models for Subtask A and B of the shared task, respectively.

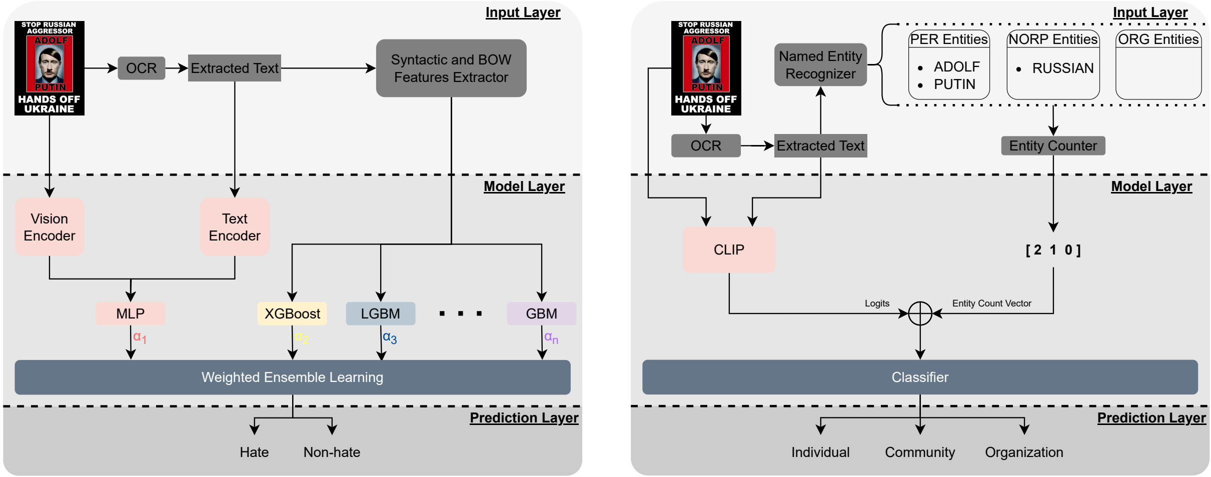
3.1 Proposed model for Subtask A: Ensemble of multimodal deep learning and text-based tabular models

The process of identifying hate speech within an image and its OCR-generated text can be approached using various methods, including relying solely on image-based or text-based models. However, in our approach, we adopt a multimodal approach to leverage the full knowledge present in the dataset. We employ both textual and visual features to train our deep learning models, aiming to capture a comprehensive understanding of the data. Additionally, we incorporate various syntactical features into our model. For this, we construct a 33-dimensional syntactical feature vector as shown in Table 2.

Furthermore, we also use the Bag-of-words (BoWs) method to extract n-grams ($n \in \{1, 2, 3\}$) from text and use them as additional features. This choice is motivated by our observation that the BoW method has competitive performance in hate speech detection and these features might possibly serve as indicators of hate speech, independent of the overall meaning conveyed by the text and image (Toraman et al., 2022b).

As illustrated in Figure 1a, our methodology begins by combining a text encoder with a vision encoder model via a multi-layer perceptron (MLP) module. This multimodal structure is initially trained on the entire training set using a linear classifier layer with the cross-entropy loss function. We select the best-performing model based on the

[1] https://github.com/therealthapa/case2023_task4
[2] https://cloud.google.com/vision/docs/ocr



(a)

(b)

Figure 1: High-level illustrations of our models for (a) Subtask A and (b) Subtask B. Each model consists of three stages, which are the Input, Model, and Prediction layers. Input layer describes the processes of text and syntactic feature extraction, and entity recognition. In Model layer, we indicate the training procedures. Furthermore, we represent the the joint learning of the models with the same colored blocks. For instance, in (a) Vision and Text encoder, and MLP is jointly trained, while XGBoost, LGBM, and GBM have independent training procedures. The last layer, i.e., Prediction, shows the classified labels for each model.

accuracy metric across multiple training epochs using the evaluation set. Subsequently, we extract the aforementioned syntactical and BoW features from the text, which are then used to train tabular learning models (i.e., classifiers), including LightGBMXT, LightGBMLarge, LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), and XGBoost (Chen and Guestrin, 2016). We then combine these models to maximize the utilization of available information. To accomplish this, we adopt an ensemble approach similar to our previous work in CASE2022 (Hürriyetoğlu et al., 2022; Sahin et al., 2022). However, this time we utilize a weighted ensembler which assigns adaptive weights to each model and generates final predictions based on these weights. The weight assignment is determined during the training phase and optimized with respect to the validation accuracy computed on the evaluation set of Subtask A.

3.2 Proposed model for Subtask B: Combining multimodal deep learning with named entity recognition

In our proposed model for Subtask B, instead of using syntactical features, we employ named entities which are extracted from the text. Named entity recognition (NER) aims to extract important information from unstructured text (Ozcelik and Toraman, 2022) and can be used as a supportive feature to improve the classification performance of a deep learning model. Therefore, we obtain named entities for the unstructured texts extracted from the text-embedded images using the spaCy library (Honnibal and Montani, 2017). SpaCy is an open source NLP library including several tasks such as Part-of-Speech (POS) tagging and NER. We use the English pretrained large NER model[3] as a named entity recognizer (see Figure 1b). The motivation behind using this model is that it contains individual, community, and organization named entity classes, which are directly related to the prediction classes of Subtask B. Therefore, we only extract PER, NORP, and ORG entities as shown in Figure 1b. The PER entities include people or fictional character names. The NORP entities represent nationalities or religious and political groups (e.g., communities). Finally, the ORG entities are referred to organization names, such as NATO.

In a previous study (Zhu, 2020), these identified entities are employed as additional textual inputs, demonstrating their contribution to the improvement of multimodal hateful meme detection. However in our work, after we obtain the aforementioned entities from the extracted texts of the images, we generate a feature vector, consisting of the counts of each entity. For instance, from Figure 1b, we represent the vector for the extracted text "STOP RUSSIAN AGRESSOR ADOLF PUTIN

[3] en_core_web_lg-3.6.0

HANDS OFF UKRAINE” as $\begin{bmatrix}2 & 1 & 0\end{bmatrix}$ since two (i.e., *Putin*, *Adolf*), one (i.e., *Russian*), and no entities are obtained for `PER`, `NORP`, and `ORG` classes, respectively.

Figure 1b shows the overall structure of our proposed model for Subtask B. Using the text-embedded images and the extracted OCR text from these images in the training set, we first fine-tune a Contrastive Language-Image Pre-Training (CLIP) model, which is a multimodal deep learning model that is pre-trained on a variety of (image, text) pairs (Radford et al., 2021). Following the completion of the CLIP training, we proceed to extract the embedding vector for each (image, text) pair in the training set of Subtask B. These embedding vectors and the entity count vector are then concatenated together to create a novel fusion vector. This newly formed vector serves as the input for training multiple tabular learning models (i.e., classifiers), including LightGBMLarge, LightGBM, and XGBoost. The classifier that achieves the highest validation accuracy score on the evaluation set of Subtask B is then selected to generate final predictions.

4 Results & Discussion

4.1 Baselines

We employ the AutoGluon framework (Erickson et al., 2020) for the implementation of our proposed models and the baselines for multimodal hate speech detection. AutoGluon is an AutoML toolkit and provides a comprehensive environment for multimodal training. We use the following hyperparameter setting for the training of all models: The learning rate is set to 1e-4, learning rate decay is set to 0.9, learning rate scheduler is cosine decay, maximum number of epochs is 10, warm-up step is 0.1, per GPU batch size is 8. During the training phase of our models and the baselines, we utilize four NVIDIA A4000 GPUs. We categorize the baselines into four categories: Tabular, Textual, Visual or Multimodal, which are explained below.

4.1.1 Tabular Baselines

For the tabular baseline models, we construct syntactic features derived from the textual data. These features, which are shown in Table 2, and BoW features (i.e., n-grams with $n \in \{1, 2, 3\}$) are employed to train classifiers including LightGBMXT, LightGBMLarge, LightGBM, CatBoost, and XGBoost. We use the AutoGluon implementation of the classifiers with default parameters.

4.1.2 Textual Baselines

For the text-only baseline models, we use the following transformer-based language models: BERT (BERT-base-cased[4]) (Devlin et al., 2018), RoBERTa (RoBERTa-base[5]) (Liu et al., 2019), DeBERTa-v3 (DeBERTa-v3-base[6]) (He et al., 2021), and ELECTRA (ELECTRA-base-discriminator[7]). We use the AutoGluon implementation of the models with a maximum token size of 512 and padding the rest.

4.1.3 Visual Baselines

For the image-only baseline models, we employ the following transformer-based encoders: Swin (swin-base-patch4-window7-224[8]), CoAtNet-v3 (coatnet-v3-rw-224-sw_in12k[9]) (Dai et al., 2021), DaViT (davit-base-msft-in1k[10]) (Ding et al., 2022), and ViT (vit-base-patch32-224-in21k[11]) (Dosovitskiy et al., 2020). We use the AutoGluon implementation of the models with default parameters.

4.1.4 Multimodal Baselines

For the multimodal baseline models where both text and images are used in the training process, we combine a textual and a visual baseline model together and jointly train them by using a multi-layer perceptron (MLP) on top of them with a binary cross-entropy loss function. To determine the optimal combination of the models, we select the top-performing text and vision encoders based on their individual performances in terms of the validation accuracy score computed on the evaluation set of the corresponding subtasks. For this, we employ the AutoGluon implementation of the text and vision encoders with a maximum token size of 512 and all other parameters set to their default values. For the classification layer, we use two fully connected linear layers (128 dimensional hidden layer) with a Leaky ReLU activation function between them. Furthermore, we also use the AutoGluon’s implementation of the CLIP model as one of the multimodal baselines.

[4] https://huggingface.co/bert-base-cased

[5] https://huggingface.co/roberta-base

[6] https://huggingface.co/microsoft/deberta-v3-base

[7] https://huggingface.co/google/electra-base-discriminator

[8] https://huggingface.co/microsoft/swin-base-patch4-window7-224-in22k

[9] https://huggingface.co/timm/coatnet_3_rw_224.sw_in12k

[10] https://huggingface.co/timm/davit_base.msft_in1k

[11] https://huggingface.co/google/vit-base-patch32-224-in21k

	Model	Precision	Recall	F1	Accuracy
Tabular	XGBoost	82.0	82.7	80.6	80.6
	LightGBM	81.2	83.5	80.3	80.4
	LightGBMLarge	81.6	82.3	80.1	80.1
	CatBoost	79.7	82.3	78.7	78.8
	LightGBMXT	78.8	81.1	77.6	77.6
Textual	ELECTRA	82.2	89.3	83.4	83.5
	BERT	79.4	84.4	79.4	79.4
	RoBERTa	84.3	81.9	81.7	81.7
	DeBERTa-v3	83.0	86.4	82.8	82.8
Visual	Swin	74.7	84.0	75.3	75.6
	CoAtNet-v3	80.4	81.1	78.8	78.8
	DaViT	81.5	79.2	78.1	78.1
	ViT	79.0	77.7	76.0	76.1
Multimodal	ELECTRA + Swin	83.3	90.1	84.5	84.6
	DeBERTa-v3 + Swin	81.8	90.9	83.8	84.0
	ELECTRA + CoAtNet-v3	85.4	86.4	84.4	84.4
	DeBERTa-v3 + CoAtNet-v3	82.9	87.6	83.2	83.3
	CLIP	79.9	91.8	82.6	82.8
Ours	**ELECTRA + Swin + Tabular**	**84.1**	**89.0**	**84.8**	**84.9**

Table 3: **Subtask A: Hate Speech Detection** evaluation results in terms of binary precision, recall, F1-score, and accuracy metrics. Tabular, textual, visual, and multimodal baselines are implemented using the AutoGluon library (Erickson et al., 2020) and categorized into their respective categories. The model which achieves the highest test scores on the final leaderboard is indicated with a bold font.

4.1.5 Our Models

For the implementation of our proposed models for Subtask A and B in Section 3, we again employ the AutoGluon library. For Subtask A, we use ELECTRA (ELECTRA-base-discriminator) and Swin (swin-base-patch4 window7-224) as our text and vision encoders, respectively. Using the syntactical and BoW features described in Section 3, we train the tabular models LightGBMXT, LightGBMLarge, LightGBM, CatBoost, and XGBoost with default parameters. Additionally, we utilize the weighted ensembler L2, an implementation provided by AutoGluon, to combine the predictions of the individual models and generate final predictions. This weighted ensembling technique assigns weights to each model, taking into account their respective classification performance on the evaluation set of Subtask A.

Furthermore, for Subtask B, we use the the multimodal baseline CLIP model and combine its embedding vector with NER features as described in Section 3. With the combined features, we train a LightGBMlarge classifier with default parameters to produce final predictions.

4.2 Evaluation Results

Table 3 and 4 show the classification performance metrics of our models and the baselines computed on the evaluation sets of Subtask A and B, respectively. *Precision*, *Recall*, *F1*, and *Accuracy* metrics are used for measuring the classification performance on the shared task of Multimodal Hate Speech Event Detection at CASE 2023[12].

The results in Table 3 and 4 clearly show that our proposed models, along with ensemble learning and using syntactical features for Subtask A and NER features for Subtask B, perform much better than all other methods, including the tabular, textual, visual, and multimodal baselines, for detecting hate speech in a multimodal setting. These results demonstrate that including different text-based features in our models improves their performance significantly, allowing us to make better use of the information in the dataset. This emphasizes the importance of using various textual attributes to enhance the overall effectiveness of the models.

In our experiments, we observe that textual methods trained with the extracted OCR text from the text-embedded images outperform visual methods trained solely on images. Additionally, the tabular models, which are trained with syntactical and BoW features (i.e., n-grams, $n \in \{1, 2, 3\}$), achieve results comparable to the text-based methods. This once again demonstrates the effectiveness of these features in multimodal hate speech detection.

Furthermore, multimodal approaches that combine multiple modalities, such as image and text, effectively leverage both textual and visual information, resulting in significantly more powerful

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

	Model	Precision	Recall	F1	Accuracy
Tabular	XGBoost	65.2	64.1	63.4	65.2
	LightGBM	68.0	67.3	66.6	68.0
	LightGBMLarge	68.8	68.3	67.4	68.8
Textual	ELECTRA	66.0	65.6	65.7	66.0
	BERT	66.0	64.7	64.7	66.0
	RoBERTa	71.7	71.4	71.4	71.7
	DeBERTa-v3	68.8	67.1	66.2	68.8
Visual	Swin	51.3	54.5	52.0	54.5
	CoAtNet-v3	49.5	50.8	49.9	50.8
	DaViT	47.9	51.6	48.5	51.6
	ViT	42.2	45.1	42.3	45.1
Multimodal	RoBERTa + CoAtNet-v3	68.5	69.6	68.4	69.6
	DeBERTa-v3 + CoAtNet-v3	63.8	63.6	62.6	63.6
	RoBERTa + Swin	72.7	73.8	72.6	73.8
	DeBERTa-v3 + Swin	66.2	66.0	65.0	66.0
	CLIP	74.2	76.8	75.4	76.8
Ours	**CLIP + NER**	**80.5**	**80.3**	**79.7**	**80.3**

Table 4: **Subtask B: Target Detection** evaluation results in terms of weighted precision, recall, F1-score, and multi-class accuracy metrics. Tabular, textual, visual, and multimodal baselines are implemented using the AutoGluon library (Erickson et al., 2020) and categorized into their respective categories. The model which achieves the highest test scores on the final leaderboard is indicated with a bold font.

Team Name	Recall	Precision	F1	Accuracy
ARC-NLP	**85.67**	**85.63**	**85.65**	**85.78**
bayesiano98	85.61	85.28	85.28	85.33
IIC_Team	85.08	84.76	84.63	84.65
DeepBlueAI	83.56	83.35	83.42	83.52
CSECU-DSG	82.52	82.44	82.48	82.62
Ometeotl	81.21	80.94	80.97	81.04
Avanthika	78.78	78.81	78.80	79.01
Sarika22	78.06	78.49	78.21	78.56
rabindra.nath	77.68	78.42	77.88	78.33
md_kashif_20	72.70	73.72	72.87	73.59
GT	52.19	52.19	52.19	52.60
Team +1	49.38	49.39	49.36	49.66
ML_Ensemblers	53.34	72.40	42.94	57.79

Table 5: The leaderboard results of **Subtask A: Hate Speech Detection**. Our team name is **ARC-NLP**. The teams are ranked by the F1 score. Our solution is ranked first in terms of all classification metrics.

Team Name	Recall	Precision	F1	Accuracy
ARC-NLP	**76.36**	**76.37**	**76.34**	**79.34**
bayesiano98	73.30	75.54	74.10	77.27
IIC_Team	68.94	71.05	69.73	72.31
Sarika22	67.77	68.41	68.05	71.49
CSECU-DSG	65.25	65.75	65.30	69.01
DeepBlueAI	64.62	66.48	65.25	69.83
Ometeotl	56.48	67.93	56.77	64.05
Avanthika	53.84	70.13	52.58	64.05
ML_Ensemblers	44.44	48.88	43.32	52.89
Team +1	34.42	35.59	33.42	35.12

Table 6: The leaderboard results of **Subtask B: Target Detection**. Our team name is **ARC-NLP**. The teams are ranked by the F1 score. Our solution is ranked first in terms of all classification metrics.

deep learning models. This integration of different modalities enhances the overall performance of the models in the process.

Finally, introducing a named entity recognition (NER) system capable of extracting key elements from unstructured text, like person names, organizations, and locations, proves particularly effective in identifying targets of hate speech (e.g., individuals, communities, and organizations) within a given text. By incorporating NER features into our model for Subtask B, we are able to further enhance the classification performance of the multimodal methods. This improvement is clearly demonstrated by the classification performance of our proposed model, as illustrated in Table 4.

4.3 Leaderboard Results

During the test phase of the shared task, we submitted our models to be evaluated on the test sets of both Subtask A and Subtask B. The test results have been presented in Table 5 and Table 6, respectively.

Our model, *ELECTRA+Swin+Tabular*, achieved the top rank among 13 participating teams in Subtask A, excelling in all classification metrics within the test results. Similarly, our model, *CLIP+NER*, secured the first position among 10 participating teams in Subtask B, performing exceptionally well across all classification metrics.

5 Conclusion

In conclusion, the utilization of text-embedded images on social media has become a common means of expressing opinions and emotions. However, it has also been exploited to spread hate speech, propaganda, and extremist ideologies, as witnessed during the Russia-Ukraine war. Detecting and addressing such instances are crucial, particularly in times of ongoing conflict. To tackle this challenge, we present our methodologies for the shared task of Multimodal Hate Speech Event Detection at CASE 2023 (Thapa et al., 2023). Our approach combines multimodal deep learning models with text-based tabular features, such as named entities and syntactical features, yielding superior performance compared to existing methods for multimodal hate speech detection. This is evidenced by achieving the first place in both Subtask A and B of the shared task on the final leaderboard, demonstrating the effectiveness of our models in identifying and categorizing hate speech events.

5.1 Ethical Considerations

This study discusses examples of harmful content (hate speech stereotypes). The authors do not support the use of harmful language, nor any of the harmful representations featured in this paper. Furthermore, the proposed models in this study are trained with the multimodal hate speech dataset described in Section 2, which specifically features the Russia-Ukraine War. Given the inherently subjective nature of the annotation process, it is reasonable to expect a certain bias towards specific subjects, individuals, organizations, and/or communities in our proposed models. We hereby acknowledge the fact that steps must be taken to mitigate this bias for future research.

5.2 Reproducibility

The multimodal hate speech dataset described in Section 2 can be accessed by contacting the authors of (Bhandari et al., 2023). Furthermore, for the reproducibility of our proposed models, we share all the necessary information such as network structure, parameter settings, libraries and tools utilized in Section 3 and 4.

References

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1993–2002.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4496–4499.

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*, pages 785–794.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. 2021. CoAtNet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977.

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*.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. 2022. DaViT: Dual attention vision transformers. In *European Conference on Computer Vision*, pages 74–92. Springer.

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*.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. Extended multilingual protest news detection - shared task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from*

Text (CASE), pages 223–228, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

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.

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 preprint arXiv:1907.11692*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Oguzhan Ozcelik and Cagri Toraman. 2022. Named entity recognition in Turkish: A comparative study with detailed error analysis. *Information Processing & Management*, 59(6):103065.

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.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.

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, Oguzhan Ozcelik, Izzet Emre Kucukkaya, and Cagri Toraman. 2022. ARC-NLP at CASE 2022 task 1: Ensemble learning for multilingual protest event detection. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 175–183.

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 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. Association for Computational Linguistics.

Cagri Toraman, Oguzhan Ozcelik, Furkan Şahinuç, and Fazli Can. 2022a. Not good times for lies: Misinformation detection on the russia-ukraine war, COVID-19, and refugees. *arXiv preprint arXiv:2210.05401*.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022b. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, Florence, Italy. Association for Computational Linguistics.

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