ARC-NLP at ClimateActivism 2024: Stance and Hate Speech Detection by Generative and Encoder Models Optimized with Tweet-Specific Elements

Ahmet Kagan Kaya and **Oguzhan Ozcelik** and **Cagri Toraman**
ASELSAN, Ankara, Turkey
kagankaya, ogozcelik, ctoraman@aselsan.com.tr

Abstract

Social media users often express hate speech towards specific targets and may either support or refuse activist movements. The automated detection of hate speech, which involves identifying both targets and stances, plays a critical role in event identification to mitigate its negative effects. In this paper, we present our methods for three subtasks of the Climate Activism Stance and Hate Event Detection Shared Task at CASE 2024. For each subtask (i) hate speech identification (ii) targets of hate speech identification (iii) stance detection, we experiment with optimized Transformer-based architectures that focus on tweet-specific features such as hashtags, URLs, and emojis. Furthermore, we investigate generative large language models, such as Llama2, using specific prompts for the first two subtasks. Our experiments demonstrate better performance of our models compared to baseline models in each subtask. Our solutions also achieve third, fourth, and first places respectively in the subtasks.

Bias Statement: This paper discusses harmful content and hate speech stereotypes. The authors do not support the use of harmful language, nor any of the harmful representations quoted below.

1 Introduction

There is a growing challenge of detecting hate speech within the context of digital communication, particularly in climate change activism, by means of natural language processing (Parihar et al., 2021). The shared task on Hate Speech and Stance Detection during Climate Activism organized in the workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) (Thapa et al., 2024) aims to provide an opportunity to study important components in identifying events during climate change activism. The task includes three subtasks for detecting (a) hate speech, (b) its target, and (c) stance being supported or opposed.

Our proposed approach in the shared task is to employ large encoder models, such as BERTweet (Nguyen et al., 2020), enhanced with Optuna (Akiba et al., 2019) to improve model performance by optimizing deep learning hyperparameters and also tweet-specific elements, such as hashtags, URLs, and emojis. Additionally, we leverage the capabilities of generative large language models, such as Llama2 (Touvron et al., 2023). Lastly, we propose hybrid solutions that benefit from both encoder and generative models. Generative models serve as a decision support mechanism, particularly in instances where the encoder model's predictions are ambiguous or uncertain.

The performances of our models are measured on the ClimaConvo dataset (Shiwakoti et al., 2024). In this study, we report the details of our solutions, which obtain 3rd place in Subtask A, 4th place in Subtask B, and 1st place in Subtask C.

2 Subtasks and Datasets

2.1 Subtasks

Subtask A: Hate Speech Detection In Subtask A, our primary objective is to develop and implement a robust hate speech detection system. In this subtask, we aim to automatically identify whether a given text contains hate speech or not, providing binary labels of "hate" and "non-hate".

Subtask B: Target Detection Subtask B aims to identify the targets of hate speech within a given hateful tweet. The dataset provided for this subtask includes labels categorizing the hate speech targets into "individual", "organization" and "community".

Subtask C: Stance Detection Subtask C aims to identify the stance in a given tweet text. The dataset provided for this subtask includes labels categorizing the stance targets into "support", "oppose", and "neutral".

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 111–117

Table 1: The distribution of the classes in train, validation, and test splits for each subtask.

Task	Class	Train	Validation	Test
A	Hate	899	190	188
	Non-Hate	6,385	1,371	1,374
B	Individual	563	120	121
	Organization	105	23	23
	Community	31	7	6
C	Support	4,328	897	921
	Oppose	700	153	141
	Neutral	2,256	511	500

Table 2: Statistics for tweet-specific elements (hashtag, URL, and emoji).

Task	Data	Avg. Htag per Tweet	Avg. URL per Tweet	Avg. Emoji per Tweet
A	Train	5.13	0.76	0.78
	Val	5.15	0.78	0.91
	Test	5.19	0.76	0.92
B	Train	7.65	0.16	0.15
	Val	7.41	0.22	0.05
	Test	7.83	0.19	0.06
C	Train	5.13	0.76	0.78
	Val	5.15	0.78	0.91
	Test	5.19	0.76	0.92

2.2 Datasets

The dataset (Shiwakoti et al., 2024) is split into train, validation, and test subsets. Table 1 gives the distribution of classes in the datasets for each subtask. The presence of hashtags, URLs, and emojis in the tweets within these datasets adds an extra layer of complexity. Table 2 presents average counts of hashtags, URLs, and emojis per tweet for each subtask. We observe that the substantial presence of hashtags, URLs, and emojis in tweets significantly impacts the predictivity of our models. These elements can be important to convey context, emotion, and additional information.

3 Main Approach

Our approach includes three solutions. First, we employ encoder models for text classification with a specific focus on tweet-specific elements such as hashtags, URLs, and emojis. Second, we employ generative large language models. Lastly, we provide hybrid solutions that benefit from both encoder and generative models. We use PyTorch (Paszke et al., 2017) and Hugging Face (Wolf et al., 2019) for model implementations.

3.1 Encoder Models

We experiment with Transformer-based architectures (Vaswani et al., 2017). The descriptions of employed models are listed below with the reasons why we select them for this task:

Megatron (Shoeybi et al., 2019): Megatron is known to perform well in hate speech detection (Toraman et al., 2022). We optimize the Megatron model in terms of the tweet features and hyperparameters using the validation dataset. The optimization process is discussed in detail in Section 3.4.

BERTweet (Nguyen et al., 2020): BERTweet has a special tokenizer that handles noisy tweet texts properly. We conduct the same optimization procedure for this model as in the Megatron model.

DeBERTa (He et al., 2021): DeBERTa shows challenging performance for text classification problems, even for noisy tweet texts (Sahin et al., 2022). We conduct the same optimization procedure for this model as in the Megatron model.

3.2 Generative Models

We employ the following open-source generative large language models. Text generation configuration has greedy decoding with a temperature setting of 1e-8 and an output length of 512 tokens.

Llama2 (Touvron et al., 2023): Llama2 is a state-of-the-art generative large language model that is specifically designed to analyze and interpret complex language patterns. This model is characterized by its large number of parameters, enabling it to process and generate highly detailed and contextually relevant text responses. We employ `Llama-2-7b-chat-hf`[1].

Mistral (Jiang et al., 2023): Mistral is an efficient model for text generation with a significantly reduced number of parameters. Its architecture not only improves computational efficiency but also detects hate speech content. We employ `Mistral-7B-Instruct-v0.1`[2] and `Mistral-7B-Instruct-v0.2`[3].

Prompts We examine existing prompts (Bach et al., 2022) to observe the performance in our preliminary experiments. We decide to use the following zero-shot prompt for Subtask A: *"Does*

[1]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[2]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[3]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Table 3: Optimized parameters of the experimented models for each subtask.

Task	Model	Hashtag Removed	URL Removed	Emoji Removed	Learning Rate	Weight Decay	Training Epoch	Training Batch	Sequence Length
A	BERTweet	✗	✗	✓	1.6e-5	0.070	3	16	128
	DeBERTa	✗	✗	✓	1.1e-5	0.027	6	16	128
	Megatron	✓	✓	✓	1.0e-5	0.010	3	16	128
B	BERTweet	✗	✗	✗	7.1e-5	0.084	9	8	128
	DeBERTa	✓	✓	✓	5.3e-5	0.049	12	8	128
C	BERTweet	✗	✗	✗	1.0e-5	0.000	3	16	96
	DeBERTa	✗	✓	✓	1.0e-5	0.000	3	16	160
	Megatron	✗	✗	✗	1.2e-5	0.035	3	8	160

this tweet convey the author's hatred towards something or someone?".

For Subtask B, we could not find existing prompts. Instead, we curate a new prompt based on our preliminary experiments: *"The goal of this subtask is to identify the targets of tweets. Give one of the labels (individual, organization, or community) for the given tweet text."*

Different from Subtask A, we observe that zero-shot prompting does not provide sufficient instruction to the model. We therefore follow few-shot prompting to provide three training examples, one for each class, in the prompt.

For Subtask C, we could not run generative models due to limited hardware and time constraints.

3.3 Hybrid Models

In Subtask A, we implement a hybrid approach that combines encoder and generative models (BERTweet+Llama2). Also, in Subtask B, we use a hybrid approach that combines encoder models and named entity recognition (BERTweet+NER).

BERTweet+Llama2 In our preliminary experiments for Subtask A, we observe that our optimized BERTweet (Nguyen et al., 2020) outperforms other encoder models. Despite its success, we observe instances where BERTweet exhibits a lack of confidence in its predictions, particularly with certain tweets that present ambiguous or subtle indications of hate speech. To address this, we incorporate Llama2 as a secondary layer of analysis. In cases where BERTweet's output logits have low confidence, i.e., lower than 0.6, we employ Llama2 to reassess the prediction label.

BERTweet+NER Following the winning model (Sahin et al., 2023) of the previous shared task (Thapa et al., 2023), we integrate named entities with the prediction output of the Transformer-based model. Named entity recognition can extract individual, organization, and community-related entities from unstructured text (Ozcelik and Toraman, 2022). We obtain entities through the spaCy library (Honnibal and Montani, 2017), employing the English Transformer pipeline model[4]. We then combine the counts of each entity with the output logits of our optimized BERTweet model. Finally, these six features are fed to a random forest model.

3.4 Optimization

We obtain our best models by optimizing the learning phase using the validation dataset. For this purpose, we employ Optuna (Akiba et al., 2019) with the following tweet-specific elements and deep learning hyperparameters:

- Hashtag: A binary feature that determines whether all hashtags are removed.
- URL: A binary feature that determines whether all URLs are removed.
- Emoji: A binary feature that determines whether all emojis are removed.
- Learning rate: Uniform range bw. 1e-5 and 1e-4.
- Weight decay: Uniform range bw. 1e-3 and 1e-1.
- Epochs: Discrete range from 3 to 10.
- Train batch size: 8, 16, or 32.
- Sequence length: 64, 96, 128, and 160

3.5 Baseline Models

We report baseline scores of four Transformer-based models provided by the organizers (Shiwakoti et al., 2024): BERT (Devlin et al., 2018), DistillBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and ClimateBERT (Webersinke et al., 2021).

4 Leaderboard Results

In this section, we report the results of all submitted models on the test data. The optimized parameters of our submitted models are reported in Table 3. Our final submitted models are listed as follows.

[4] en_core_web_trf

Table 4: **Subtask A: Hate Speech Detection.** Test results in terms of precision, recall, F1 score, and accuracy. The model which achieves the highest test scores on the final leaderboard is indicated with a bold font. Baseline scores are obtained from Shiwakoti et al. (2024).

	Model	Pre	Rec	F1	Acc
Baseline	BERT	-	-	0.7080	0.9010
	DistilBERT	-	-	0.6640	0.8960
	RoBERTa	-	-	0.6620	0.8420
	ClimateBERT	-	-	0.7040	0.8840
Ours	Megatron	0.8003	0.9415	0.8532	0.9475
	BERTweet	0.8687	0.8923	0.8800	0.9507
	DeBERTa	0.8623	0.8836	0.8725	0.9475
	Llama2	0.5248	0.3894	0.4471	0.8827
	Mistralv0.1	0.5416	0.1368	0.2184	0.8808
	Mistralv0.2	0.3571	0.3947	0.3750	0.8398
	BERTweet+Llama2	0.8973	0.8833	**0.8901**	0.9526

Table 5: **Subtask B: Target Detection.** Notations are the same as Table 4.

	Model	Pre	Rec	F1	Acc
Baseline	BERT	-	-	0.5540	0.6410
	DistilBERT	-	-	0.5500	0.6030
	RoBERTa	-	-	0.5010	0.7160
	ClimateBERT	-	-	0.5490	0.6040
Ours	BERTweet	0.7728	0.7588	**0.7638**	0.9133
	DeBERTa	0.7149	0.7005	0.6997	0.9000
	BERTweet+NER	0.7421	0.7588	0.7500	0.9133
	DeBERTa+NER	0.7149	0.7005	0.6997	0.9000
	Llama2	0.5775	0.5152	0.4439	0.8067

Table 6: **Subtask C: Stance Detection.** Notations are the same as Table 4.

	Model	Pre	Rec	F1	Acc
Baseline	BERT	-	-	0.4660	0.5860
	DistilBERT	-	-	0.5270	0.6100
	RoBERTa	-	-	0.5420	0.6480
	ClimateBERT	-	-	0.5450	0.6510
Ours	Megatron	0.7509	0.7200	0.7342	0.7298
	BERTweet	0.7848	0.7226	**0.7483**	0.7490
	DeBERTa	0.7555	0.7242	0.7385	0.7356

Subtask A Our hybrid model (BERTweet+ Llama2) gets the 3rd place among 22 participants.

Subtask B Our optimized encoder (BERTweet) gets the 4th place among 18 participants.

Subtask C Our optimized encoder (BERTweet) gets the 1st place among 19 participants.

In Table 4, we present evaluation results for Subtask A, highlighting the better performance of our optimized BERTweet model, particularly over DeBERTa. This might show that the special tweet tokenizer can handle noisy tweet text. Generative models, Llama2 and Mistral, misinterpret some tweets (e.g., the tweets having many hashtags). We obtain better performance when they are used as a support tool for BERTWeet in uncertain cases.

In Table 5, we report that the optimized BERTweet model outperforms others in Subtask B, while the inclusion of named entities does not enhance performance for identifying individual, organization, and community targets. This ineffectiveness can be attributed to the prevalence of "individual" entities such as Greta Thunberg surpassing other entities. Moreover, Llama2 performs poorly using few-shot prompts. Unlike Subtask A, we do not integrate Llama2 with BERTweet, since output logits are mostly above the confidence threshold.

In Table 6, we report the evaluation results for Subtask C. We obtain our highest score by using an optimized version of BERTweet. It has a short length of input tokenization (96 tokens) with special tokens for tweet-specific elements. We could not implement generative models for Subtask C due to limited hardware and time constraints. Nevertheless, we obtain the highest score among other participants in this subtask.

5 Discussion

Encoder Models and Tweet-specific Elements We observe that the removal of hashtags and URLs is insignificant for the BERTweet model. This is because the it demonstrates proficient tokenization even when hashtags and URLs are retained. However, in Subtask A, the removal of emojis is an important hyperparameter since we observe that emojis can be misleading in detecting hate speech.

Generative Models and Error Analysis We provide sample outputs generated by Llama2 with zero-shot prompts for Subtask A in Table 7. We find that there are many near-duplicate tweets that have many hashtags but similar content regarding Greta Thunberg (e.g. False Negative in the table). Most of them are predicted as not hateful by the generative model, however the ground truth is labeled as hate. We suspect that the high number of such false negatives profoundly could impact its overall accuracy. On the other hand, the generative model gives a reasonable response such that the tweet content is critical of Greta Thunberg, instead of hate speech. The degree of offensiveness in this tweet is indeed controversial. This may show that the knowledge of generative models does not necessarily align with human annotation for hate speech detection.

Table 7: Error analysis of Llama2 with zero-shot prompts for Subtask A Hate Speech Detection. TP: True Positive, FN: False Negative, FP: False Positive, TN: True Negative.

	Tweet	Ground Truth	Response
TP	#India least bothered over #ClimateCrisis. Majority argues they are preoccupied with their own devised Hindu-Muslim issues. Irony, #policymakers hiding the fact: #climate disaster will wipe everyone out along with progeny, or the threat is ignored? #FridaysForFuture byadavbjp https://t.co/2LAESz7s9Y	Hate	Yes. The author of the tweet is expressing frustration and disappointment towards the people of India for being more concerned about their own religious differences (Hindu-Muslim issues) than taking action to address the climate crisis, which they believe will have severe consequences for everyone, including their own progeny
FN	#ExtinctionRebellion #ClimateCrisis #ClimateAction #GlobalWarming #FridaysForFuture #ClimateChange #Greenwashing #Renewables #ClimateStrike You've been fooled by Greta Thunberg:,	Hate	No, the tweet does not convey the author's hatred towards Greta Thunberg or anyone else. The tweet is critical of Greta Thunberg and the movement she is leading, but it does not express hatred towards her.
FP	•Week 19• The Time Is Ticking,And We Have No Chance For Saying, "We'll Do This/That Tomorrow,For The Environment,ACT NOW!" #ActNow #FridaysForFuture #ClimateActionNow @Fridays4future @FFFMAPA @Riseupmovt @UNEP @spicewarriors https://t.co/IWs0cIYTpn,	Non-Hate	Yes. The tweet conveys a sense of urgency and frustration towards the lack of action being taken to address the climate crisis, with the use of phrases like "We have no chance" and "ACT NOW!"
TN	Let us celebrate the PLANET and take care of EARTH #Climatechange #ClimateStrike #ClimateAction #ClimateCrisis #savetheearth #savetheplanet #FridaysForFuture #Bangladesh https://t.co/xrwYnQuDJw,	Non-Hate	No, the tweet does not convey the author's hatred towards something or someone. The tweet is focused on celebrating the planet and taking care of the Earth, using hashtags related to climate change and environmental activism. There is no indication of hatred or negative sentiment towards any particular entity.

We provide another sample for a False Positive prediction. The model response is non-trivial such that it gives some examples of frustration in the tweet. However, human annotation is not hateful for this tweet. This is also another example of an alignment problem between the knowledge of generative models and human annotations for climate activism and hate speech detection.

In Table 7, we also provide a sample case where our hybrid solution, BERTweet+Llama2, is useful in this task. The True Positive (TP) sample in the table is predicted as non-hate by BERTweet with a confidence score of 0.6. However, Llama2 evaluates this tweet as hate with an insightful explanation.

6 Conclusion

We conclude that the optimized BERTweet model outperforms other encoder models in all subtasks, indicating the importance of tweet-specific elements (hashtag, URL, and emoji) in hate event detection. Overall, generative models perform poorly in this task. More investigation is needed to understand their capabilities for hate speech detection. A possible reason for poor performance could be our prompts or generation config. Nevertheless, the support of Llama2 increases the performance in Subtask A.

In future work, state-of-the-art generative models like GPT3.5[5] or GPT4[6] can be employed in addition to Llama2 and Mistral. Moreover, prompt tuning can improve the performance of generative models and extend the work for generalizing model understanding capacity.

7 Limitations

The dataset has only English text in this study. More experiments in different languages can be conducted to generalize the results to other languages. Also, the optimized hyperparameters for encoder models are limited to the dataset used in this study. Generative models may in some instances produce inaccurate, biased, or other objectionable responses to user prompts.

8 Ethics Statement

The authors do not support the use of harmful language or any of the harmful representations featured in this paper. Furthermore, our proposed models are trained on an annotated dataset; therefore, they may have certain bias towards specific subjects, individuals, organizations, and communities. We acknowledge the necessity of bias mitigation for future research. Lastly, for reproducibility, we share details such as hyperparameters, libraries, and tools in Section 3, and the datasets are published by Shiwakoti et al. (2024).

[5]https://platform.openai.com/docs/models/gpt-3-5
[6]https://openai.com/gpt-4

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

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*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining 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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 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.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

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.

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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. *Preprint*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

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, Kritesh Rauniyar, Farhan Ahmad Jafri, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024. Stance and hate event detection in tweets related to climate activism - 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)*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

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*.