AAST-NLP at ClimateActivism 2024: Ensemble-Based Climate Activism Stance and Hate Speech Detection : Leveraging Pretrained Language Models

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

Abstract

Climate activism has emerged as a powerful force in addressing the urgent challenges posed by climate change. Individuals and organizations passionate about environmental issues use platforms like Twitter to mobilize support, share information, and advocate for policy changes. Unfortunately, amidst the passionate discussions, there has been an unfortunate rise in the prevalence of hate speech on the platform. Some users resort to personal attacks and divisive language, undermining the constructive efforts of climate activists. In this paper, we describe our approaches for three subtasks of ClimateActivism at CASE 2024. For all the three subtasks, we utilize pretrained language models enhanced by ensemble learning. Regarding the second subtask, dedicated to target detection, we experimented with incorporating Named Entity Recognition in the pipeline. Additionally, our models secure the second, third and fifth ranks in the three subtasks respectively.

1 Introduction

Climate activism has emerged as a formidable force in contemporary society, reflecting a collective global consciousness towards environmental stewardship. The advocates of climate activism ardently emphasize the urgency of addressing climate change as a paramount global challenge. Through various channels, such as organized protests, advocacy campaigns, and international collaborations, climate activists strive to raise awareness about the detrimental impact of human activities on the planet's ecosystems (Fisher and Nasrin, 2020). Social media has played a pivotal role in amplifying the voices of climate activists, providing a powerful platform for the dissemination of information and the mobilization of global communities. Platforms like Twitter, Instagram, and Facebook have facilitated the rapid spread of awareness campaigns, enabling activists to reach diverse audiences and garner widespread support for climate action.(Arnot et al., 2024; Gómez-Casillas and Márquez, 2023) However, the same social media channels have also been susceptible to the spread of misinformation and targeted attacks against climate activists (Levantesi). Instances of hate speech and online harassment have, unfortunately, been prevalent, underscoring the double-edged nature of social media in the context of climate activism. The Climate Activism 2024 shared task (Thapa et al., 2024) delves into this significant subject by providing a dataset that encourages collaboration among researchers to address this crucial issue. The paper is organized into several key sections: related work, dataset and task description, methodology, results, and a discussion leading to a conclusion.

2 Related Work

In the realm of social media, the challenge of hate speech detection arises as a pressing concern (Jahan and Oussalah, 2023b). A number of researcher have proposed models to tackle this issue. Language models, in particular, have been a major driving force or this recent succes. Roberta, for instance, was used in detecting hate speech from social media data (Alonso et al., 2020). Some BERT based models were trained specifically for hate speech detection and achieved incredible results (Caselli et al., 2021). Language models were also adapted to multiple languages and were noticed to perform high results (Mujahid et al., 2023; Plaza-Del-Arco et al., 2021). A number of papers provide a comprehensive overview over the latest challenges and trend in hate speech detection, some of which serve as a starting point for any researcher working on this topic (Parihar et al., 2021; Jahan and Oussalah, 2023a). Hate speech manifests in various forms, and scholars have focused on creating systems to tackle issues like Cyber Bullying (Akhter et al., 2023; Hsien et al., 2022), racism (Schütz et al., 2021), and sexism (Plaza et al., 2023).

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 105–110

Despite the ongoing and comprehensive endeavors of researchers, as far as we are aware, there has not been a unified research initiative to monitor hate speech specifically directed at climate activists, a significant and alarming occurrence.

3 Dataset & Task

The shared task on Climate Activism Stance and Hate Event Detection at CASE 2024[1]consists of three main subtasks.Each subtask will be discussed in details in the following subsections. The provided dataset primarily comprises tweets expressing either support or opposition towards climate activists in various contexts (Shiwakoti et al., 2024). The subsequent subsections will present an overview of the distribution for each dataset, emphasizing the challenges posed by imbalances, particularly instances where certain classes were underrepresented.

3.1 Subtask A: Hate Speech Detection

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

	Training	Validation	Testing
No Hate	6385	1371	1374
Hate	899	190	188
Overall	7284	1561	1562

Table 1: Subtask A’s Dataset Distribution.

3.2 Subtask B: Targets of Hate Speech Identification

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

	Training	Validation	Testing
Individual	563	120	121
Organization	105	23	23
Community	31	7	6
Overall	699	150	150

Table 2: Subtask B’s Dataset Distribution.

[1]https://codalab.lisn.upsaclay.fr/competitions/16206

3.3 Subtask C: Stance Detection

The third subtask is a multiclass classification problem where tweets given are classified into three distinct classes: "Support", "Oppose", and "Neutral". Table 3 illustrates the data distribution for the different classes within the dataset.

	Training	Validation	Testing
Support	4328	897	921
Oppose	2256	153	141
Neutral	700	511	500
Overall	7284	1561	1562

Table 3: Subtask C’s Dataset Distribution.

3.4 Data Preprocessing

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

- Removal of punctuation as many tweets contained .
- Applying PySpellChecker[2] to check for misspelled words and correct them.
- Removal of hyperlinks and emojis as they did not meaning needed for our classification process.
- Removal of hashtags and tags as most of the text contained relatively similar hashtags like #ClimateChange and #ClimateStrike.

4 Methodology

In the following subsections, we will expand on the proposed models for each subtask. We will also expand on the main ideas we experimented on to tackle the class imbalance issue we encountered.

4.1 Proposed Model

4.1.1 Language Models

Several language models were experimented with through the process of fine-tuning, driven by their

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

remarkable performance in the context of our specific topic, We finetuned RoBERTa(Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020) and HateBERT (Caselli et al., 2021) on all of the datasets. Roberta showed superior performacne in terms of f1-score on all of the subtasks as will be shown in the results section5. However, XLM-RoBERTa and HateBERT where shown to shine in different aspects either achieving higher recall or precision, something that encouraged us to use our ensemble-based approach.

4.1.2 NER Based Classifier

For Subtask B, we experimented with 2 Named Entity Recognition modules, SpaCy[3] and a BERT based NER[4], to extract important landmarks. The BERT based NER showed superior performance in extracting names whilst SpaCy was used to extract ORG and NoORG landmarks. This approach was inspired by (Sahin et al., 2023) work on multimodal hate speech detection. The extracted features would then be classified using a classifier or simply through checking which token appeared the most and assigning the class accordingly. To further illustrate how the NER works, consider the following dataset sample after it went through preprocessing "You ve been fooled by Greta Thunberg" the NER would report the following tokens illustrated in Table 4.

Class	Person	ORG	NoORG
Token Count	1	0	0

Table 4: NER Tokens extracted.

4.2 Ensembling

Ensembling machine learning models involves combining diverse models to improve robustness, generalization, and predictive performance. Our strategy employs hard voting, where individual models within the ensemble make predictions on a dataset, and the final prediction is determined by majority voting. We conducted experiments involving the ensemble of top-k learners for each subtask, culminating in the derivation of our predictions.

4.3 Tackling Class Imbalance

4.3.1 Resampling

Resampling involves modifying the distribution of training datasets to elevate the significance of minority classes (Kraiem et al., 2021), Random under-sampling (RUS) entails randomly removing data points from the majority class, while random oversampling (ROS) involves duplicating instances from the minority class. Both ROS and RUS were employed to address the imbalance in the dataset, yet ROS was the one incorporated in the final submission as it was found to increase the f1-score.

4.3.2 Loss Functions

Several loss functions were experimented with, and initially, Weighted Cross-Entropy loss was employed for our subtasks. The weights were calculated using the scikit[5] class weight function, resulting in a slight improvement. Focal Loss was also used yet it provided us with minimal improvements. Ultimately, an experiment was conducted using Dice Loss, a customized loss function tailored to NLP tasks based on the Sørensen–Dice coefficient (Li et al., 2019).

4.4 Experiment Settings

The training procedure was conducted using the Google Colab [6] platform for training our pipeline, which has 12.68 GB of RAM, a 14.75 GB NVIDIA Tesla T4 GPU, and Python language.We employed the autofit functionality from ktrain (Maiya, 2022), incorporating a triangular learning rate policy (Smith, 2017). The specific parameters chosen for our experiment are outlined in the table below.

Hyperparameter	Value
Epochs	30
Learning Rate	2e-5
Batch Size	16
Max length	40
Optimizer	Adam
Early Stopping Patience	5
Reduce On Plateau	2
Loss Function	Dice Loss

Table 5: Training Hyperparameters.

5 Results

This section elaborates on the results obtained from using the mentioned systems. It's crucial to note that RoBERTa, XLM-RoBERTa, and HateBERT underwent multiple training sessions with varying

[3] https://spacy.io/
[4] https://huggingface.co/dslim/bert-base-NER
[5] https://scikit-learn.org/stable/
[6] https://colab.google/

dataset distributions through resampling. Additionally, both base and large versions were experimented with for RoBERTa and XLM-RoBERTa. The Top-k Ensemble method selected the highest k submissions for ensembling.

5.1 Subtask A

Table 6 provides a visual representation of how the mentioned models performed on the test set. It is evident that certain models outperformed others in specific metrics. Notably, Roberta achieved the highest precision among all models, while HateBERT exhibited the highest recall among the reported models. These findings prompted us to adapt our ensemble approach, aiming to leverage the strengths of various models.

Model	Precision	Recall	F1-Score
RoBERTa	**0.8688**	0.8775	0.8731
XLM-RoBERTa	0.8544	0.9174	0.8824
HateBERT	0.7994	**0.9611**	0.8579
Top-3 Ensemble	0.8544	0.9174	0.8824
Top-5 Ensemble	0.8654	0.9231	**0.8914**

Table 6: Results For Subtask A.

5.2 Subtask B

Table 7 illustrates the performance of the previously mentioned models on the test set. Roberta significantly surpasses the performance of all other models, with XLM-RoBERTa also demonstrating relatively strong performance.The NER-based classifier exhibited solid performance, even outperforming HateBERT. Employing a hard voting scheme to ensemble predictions, with greater emphasis on RoBERTa, resulted in consistently high outcomes.

Model	Precision	Recall	F1-Score
RoBERTa	0.7416	0.7501	0.7434
XLM-RoBERTa	0.7271	0.7194	0.7232
HateBERT	0.7071	0.6788	0.6919
NER Based	0.7123	0.7185	0.7063
Top-3 Ensemble	0.7561	0.7629	0.7570
Top-5 Ensemble	**0.7706**	**0.7689**	**0.7665**

Table 7: Results For Subtask B.

5.3 Subtask C

Table 8 illustrates the performance of the previously mentioned models on the test set. Roberta slightly surpassed the other two models in performance. However, upon ensembling the three models, we observed only a slight improvement in performance. This raises a pertinent question about whether the marginal increase, in our specific case, justifies the computational costs associated with real-time implementation for this subtask.

Model	Precision	Recall	F1-Score
RoBERTa	**0.7169**	0.7664	0.7356
XLM-RoBERTa	0.7022	0.7154	0.7070
HateBERT	0.7001	0.7869	0.7319
Top-3 Ensemble	0.7078	**0.7931**	**0.7398**

Table 8: Results For Subtask C.

5.4 Leaderboard Results

During the evaluation phase of the shared task, we submitted our models for assessment on the test sets of both Subtask A, Subtask B and Subtask C. The outcomes of the tests are presented in Table 6, Table 7 and Table 8, respectively. Our ensemble based approach, which combines multiple BERT-based models, achieved the second place among the 23 participating teams in Subtask A. Similarly, the same model secured the second position among the 18 participating teams in Subtask B. Whilst in subtask C, our model achieves the fifth place.

6 Discussion & Future Work

The results obtained show that leveraging pre-trained models for the classification of hate tweets could provide very promising results, even when faced with unbalanced data. These results form a great basis for further research, including but not limited to incorporating more language models into the ensemble, such as the FALCON series of models (Almazrouei et al., 2023) or Mistral (Jiang et al., 2023). Creating synthetic data with the aim of enhancing model robustness or improving performance on underrepresented classes or ones the model faces difficulties in identifying is also an intriguing strategy. Attempting different hyperparameter configurations is also worthy of further investigation. Overall, with further refinement, this approach could definitely have a real impact on

reducing the hate experienced by climate activists all around the world.

7 Conclusion

This study centers on analyzing tweets that convey opinions and emotions, but regrettably, these tweets are also employed as channels for disseminating hate speech, propaganda, and extremist ideologies. Particularly, amidst the recent surge in climate activism, social media emerged as a primary platform not just for raising awareness but unfortunately for spreading negativity as well. The increasing prevalence of offensive content on social media presents challenges in efficiently identifying and moderating such material. To tackle this alarming issue, we present our solution based on ensembling top-k performing models. Language models remain the crucial tool for addressing contemporary Natural Language Processing (NLP) challenges, consistently attaining top positions across various subtasks. Our research findings paves the way for upcoming enhancements to address and mitigate this highly concerning issue in the near future.

References

Arnisha Akhter, Uzzal Kumar Acharjee, Md. Alamin Talukder, Md. Manowarul Islam, and Md. Ashraf Uddin. 2023. A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Natural Language Processing Journal*, 4:100027.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Pedro Alonso, Rajkumar Saini, and György Kovacs. 2020. TheNorth at SemEval-2020 task 12: Hate speech detection using RoBERTa. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2197–2202, Barcelona (online). International Committee for Computational Linguistics.

Grace Arnot, Hannah Pitt, Simone McCarthy, Chloe Cordedda, Sarah Marko, and Samantha L. Thomas. 2024. Australian youth perspectives on the role of social media in climate action. *Australian and New Zealand Journal of Public Health*, 48(1):100111.

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

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

Dana R. Fisher and Sohana Nasrin. 2020. Climate activism and its effects. *WIREs Climate Change*, 12(1).

Amalia Gómez-Casillas and Victoria Gómez Márquez. 2023. The effect of social network sites usage in climate change awareness in Latin America. *Population and Environment*, 45(2).

Yeo Khang Hsien, Zailan Arabee Abdul Salam, and Vinothini Kasinathan. 2022. Cyber Bullying Detection using Natural Language Processing (NLP) and Text Analytics. *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*.

Md Saroar Jahan and Mourad Oussalah. 2023a. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Saroar Jahan and Mourad Oussalah. 2023b. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mohamed S. Kraiem, F. Sánchez, and María N. Moreno García. 2021. Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. an approach based on association models. *Applied sciences*, 11(18):8546.

Stella Levantesi. "Enemies of Society": How the media portray climate activists.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced NLP tasks. *arXiv (Cornell University)*.

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.

Arun S. Maiya. 2022. ktrain: A low-code library for augmented machine learning.

Muhammad Mujahid, Khadija Kanwal, Furqan Rustam, Wajdi Aljadani, and Imran Ashraf. 2023. Arabic ChatGPT tweets classification using ROBERTA and BERT ensemble model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–23.

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.

Laura Plaza, Jorge Carrillo-De-Albornoz, Roser Morante, Enrique Amigó, Julio A. Gonzalo, Damiano Spina, and Paolo Rosso. 2023. 68 *Overview of EXIST 2023: SEXism Identification in Social NETworks*.

Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín–Valdivia. 2021. Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

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.

Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepčević, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Strobel, and Matthias Zeppelzauer. 2021. Automatic Sexism Detection with Multilingual Transformer Models. *arXiv (Cornell University)*.

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*.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks.

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)*.