Z-AGI Labs at ClimateActivism 2024: Stance and Hate Event Detection on Social Media

Nikhil Narayan
Z-AGI Labs
nikhilnarayan73@gmail.com

Mrutyunjay Biswal
Z-AGI Labs
mrutyunjay.biswal.hmu@gmail.com

Abstract

In the digital realm, rich data serves as a crucial source of insights into the complexities of social, political, and economic landscapes. Addressing the growing need for high-quality information on events and the imperative to combat hate speech, this research led to the establishment of the Shared Task on Climate Activism Stance and Hate Event Detection at CASE 2024. Focused on climate activists contending with hate speech on social media, our study contributes to hate speech identification from tweets. Analyzing three sub-tasks - Hate Speech Detection (Sub-task A), Targets of Hate Speech Identification (Sub-task B), and Stance Detection (Sub-task C) - Team Z-AGI Labs evaluated various models, including LSTM, Xgboost, and LGBM based on Tf-Idf. Results unveiled intriguing variations, with Catboost excelling in Subtask-B (F1: 0.5604) and Subtask-C (F1: 0.7081), while LGBM emerged as the top-performing model for Subtask-A (F1: 0.8684). This research provides valuable insights into the suitability of classical machine learning models for climate hate speech and stance detection, aiding informed model selection for robust mechanisms.

1 Introduction

In the ever-evolving landscape of our digital era, an expansive tapestry of data unfolds, revealing profound insights into the intricate dynamics of social, political, and economic systems. The narratives of citizen responses to COVID policies (2020-2022) and the unfolding Russia-Ukraine conflict stand out as crucial chapters(Tanev et al., 2023), vividly demonstrating the indispensable role of event-centric data in unraveling the multifaceted tapestry of real-world scenarios. These narratives underscore the pressing need for sophisticated tools capable of discerning and addressing hate speech, ultimately leading to the inception of the Shared Task on Climate Activism Stance and Hate Event Detection at CASE 2024(Thapa et al., 2024).

Within the realms of social media platforms, where climate activists converge to share insights, mobilize support, and voice concerns, instances of hate speech can emerge, casting a shadow over the collaborative spirit of the movement. Sub-task A of our shared task(Shiwakoti et al., 2024), Hate Speech Detection, emerges from the very fabric of these real-world scenarios, challenging participants to meticulously scrutinize textual content for the presence of hate speech. Navigating the landscape of hate speech requires a profound understanding of its targets. Real-world examples abound, illustrating instances where individual activists, environmental organizations, and entire communities face the brunt of hateful rhetoric. Sub-task B, Targets of Hate Speech Identification, mirrors these authentic situations by urging participants to categorize hate speech targets into "individuals," "organisations," or "communities." In witnessing the unfolding narratives of climate activism, the importance of understanding stance dynamics becomes evident. Real-world scenarios often involve a spectrum of sentiments — from unwavering support to vehement opposition or maintaining a neutral stance. Sub-task C, Stance Detection, captures the essence of these dynamic narratives, prompting participants to decipher the sentiments expressed in textual content. By doing so, participants contribute to a deeper understanding of how the collective sentiment shapes the discourse surrounding climate change events.

The shared task thus emerges not as a detached academic exercise(Parihar et al., 2021) but as a direct response to the challenges faced in the trenches of climate activism. Through real-world instances and tangible connections, participants are invited to be catalysts for positive change, developing tools that align with the authentic dynamics of the digital discourse in climate change activism. In this endeavor, the shared task serves as a bridge between the virtual and the real, fostering a more resilient

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 161–165
March 22, 2024 ©2024 Association for Computational Linguistics

and empathetic space for those advocating for a sustainable and equitable future.

In this paper, we describe our approach to tackle the challenges. From here, the report continues in the following manner: In section 2, we give an overview of the dataset for each subtask and describe the challenge at hand. In section 3, we present our approach in detail, covering the intricacies of our experimental set-up, cross-validation strategy, models used, and intuition behind them. In section 4, we brief the results from the experiments section. Then, we conclude in section 5 with the final takeaways, our standings, and the scope of future work.

2 Dataset Description

This section provides an overview of the dataset designed to facilitate the exploration and evaluation of these key aspects.

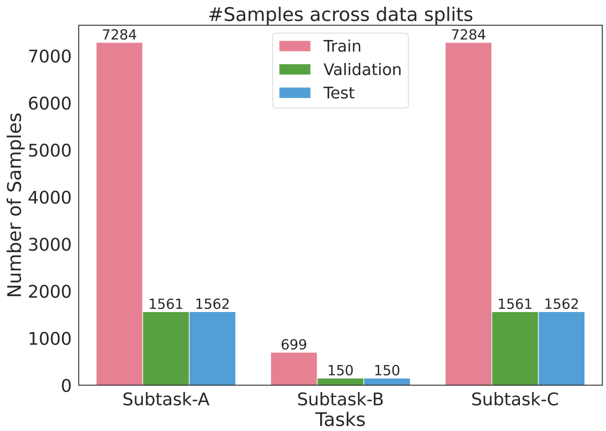

Figure 1: Train-Val-Test Split for different subtasks.

2.1 Hate Speech Detection (Sub-task A)

The primary objective of Sub-task A is to determine the presence or absence of hate speech within a given text. The text dataset for Sub-task A is enriched with binary annotations, explicitly indicating the prevalence of hate speech. Each instance is marked to signify whether it contains hate speech or remains devoid of such content. The Dataset provided for the task contains 7284 samples in the train set, 1561 samples in the Validation set and 1562 samples in the test set.

2.2 Targets of Hate Speech Detection (Sub-task B)

Sub-task B is dedicated to identifying the specific targets of hate speech within hateful texts. The

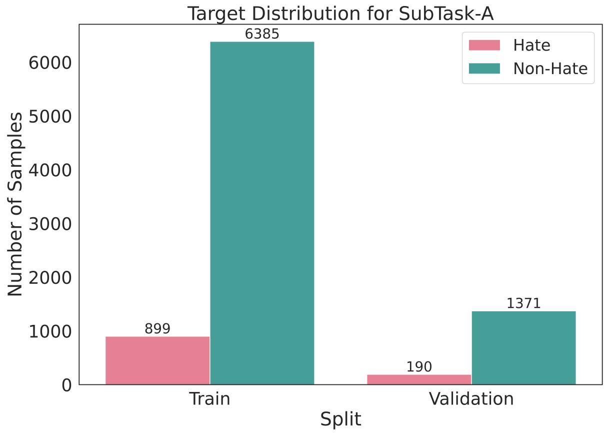

Figure 2: Target Distribution for Subtask-A.

dataset for Sub-task B is meticulously annotated to delineate the entities targeted by hate speech. Annotations classify the targets into three distinct categories: "individual," "organization," and "community." The Dataset provided for the task contains 699 samples in the train set, and 150 samples in both the Validation set and the test set.

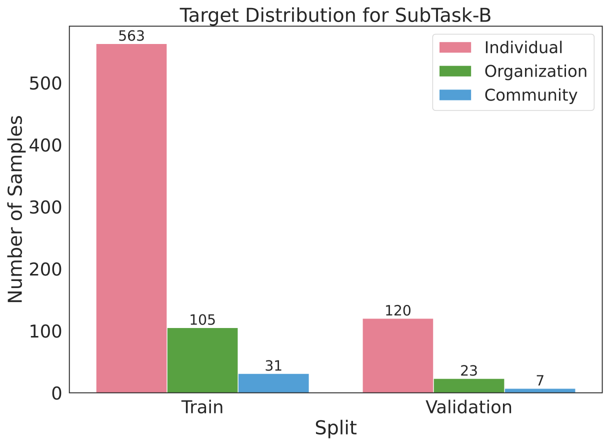

Figure 3: Target Distribution for Subtask-B.

2.3 Stance Detection (Sub-task C)

Sub-task C focuses on discerning the stance expressed in a given text within the context of climate change activism. The text dataset for Sub-task C is annotated to capture three distinct stances: "support," "oppose," and "neutral." The Dataset provided for the task contains 7284 samples in the train set, 1561 samples in the Validation set and 1562 samples in the test set.

3 Experimental Set-Up

In this section, we delve into our methodology and the specifics of the experimental setup. For

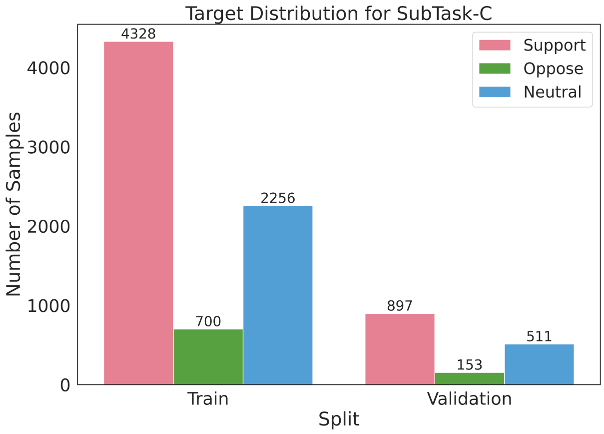

Figure 4: Target Distribution for Subtask-C.

each dataset, we first develop a validation technique. Since every dataset is not fairly balanced, we choose to use Stratified K-Fold cross-validation with 5 folds. Additionally, we used 42 as the random seed when generating the splits.

3.1 Preprocessing

The preprocessing phase plays a pivotal role in refining the content for subsequent feature extraction. Upon careful examination, it was noted that the majority of tweets exhibit a notable absence of emojis or redundant punctuation marks that necessitate attention. Although a substantial portion of the content is successfully cleansed, a distinctive characteristic emerged: the prevalence of extensive hashtags across all tweets. Furthermore, a noteworthy observation was made regarding tweets with similar textual content but distinct hashtags, resulting in disparate outcomes. To address these intricacies, the preprocessing pipeline involves the removal of URLs and hyperlinks associated with the content. Specifically, the focus is directed towards the hashtags, which undergo further processing using the Ekphrasis(Baziotis et al., 2017) tokenizer to segment them into semantically meaningful tokens. Notably, the decision was made to employ the Tokenizer Separator token to distinctively segregate normal text from hashtag texts. In the case of the former, tweet preprocessor was applied to facilitate the cleansing process.

3.2 Modeling

Our methodology commences with the establishment of baseline scores using Tf-Idf in conjunction with Naive Bayes for each of the three subtasks. This initial step allows us to gauge the performance of a rudimentary model before advancing to more sophisticated approaches. Moving beyond the baseline, we employ powerful classical machine learning models, namely Random Forest, Xgboost(Chen and Guestrin, 2016), CatBoost(Prokhorenkova et al., 2018), and LGBM(Ke et al., 2017), leveraging Tf-Idf as the feature extraction method. This ensemble of classical models provides a comprehensive understanding of the task's intricacies and sets a benchmark for further exploration. We also used hyperparameter tuning using optuna for models like Xgboost, CatBoost and LGBM.

To delve into the nuances of textual content and capture intricate dependencies, we introduce a deep learning approach. Our model architecture encompasses a bi-directional LSTM-based(Sundermeyer et al., 2014) framework with attention mechanisms(Vaswani et al., 2017). Specifically, two bi-directional LSTM layers precede an attention block, enhancing the model's capacity to grasp sequential patterns. The attention head is intricately connected through two dense layers, culminating in a sigmoid activation function in the final layer. The model is trained using the Adam optimizer(Kingma and Ba, 2014) and Binary Cross Entropy as the loss function. Crucial hyperparameters, including batch size, number of epochs, learning rate, vocabulary size, embedding dimension, and maximum length of the input sequence, undergo meticulous tuning on a case-to-case basis to optimize model performance.

We leverage the capabilities of Transformer-based language models to improve downstream job performance, taking into account the small sample size of the available datasets. These models use fine-tuning on the encoder layers while keeping the embedding layers frozen to maintain contextual knowledge that has already been learned. TFAutoModelForSequenceClassification is adopted as the Transformer-based model, with corresponding hyperparameters tailored for each subtask.

To ensure computational efficiency and scalability, all training and inference operations are carried out using the Kaggle runtime, Google Colab, and a MacBook Pro M1 with 16GB of unified memory.

4 Results

All the subtasks were evaluated using F1 Score, Precision, Recall, and Accuracy. It is evident from the results matrix 1 that the LSTM based model poses a strong competition in performance for all the subtasks nearing the best score for all the sub-

Models	Subtask-A	Subtask-B	Subtask-C
LSTM + Attention	0.8433	0.5370	0.5008
Tf-Idf + Logistic Regression	0.8516	0.5577	0.7075
Tf-Idf + LGBM	**0.8684**	0.5097	0.6055
Tf-Idf + CatBoost	0.8586	**0.5604**	**0.7081**
Tf-Idf + Xgboost	0.8228	0.5360	0.6994
Tf-Idf + Random Forest	0.8548	0.5496	0.6765
Tf-Idf + Naive Bayes	0.8516	0.5482	0.6065

Table 1: F1-Scores of different approaches

Team	Precision	F1-Score	Accuracy	Recall
mrutyunjay_research	**0.9686 (1)**	0.8539 (15)	0.9494 (6)	0.7922 (19)
refaat1731	0.9607 (2)	0.8556 (12)	0.9494 (6)	0.7968 (18)
kagankaya1	0.9415 (3)	0.8532 (16)	0.9475 (8)	0.8003 (17)
htanev	0.9246 (4)	0.8310 (18)	0.9405 (13)	0.7779 (20)
kojiro000	0.9226 (5)	**0.8699 (7)**	**0.9507 (5)**	**0.8319 (14)**

Table 2: Snippet of Leaderboard sorted by Recall for SubTask-1

Username	Recall	Precision	F1-Score	Accuracy
AhmedElSayed	0.7078 (8)	0.7931 (1)	0.7398 (6)	0.7439 (4)
mrutyunjay_research	0.6294 (16)	0.7926 (2)	0.6372 (16)	0.6908 (12)
gh_mhdi	0.7145 (5)	0.7863 (3)	0.7447 (4)	0.7311 (8)
kagankaya1	**0.7226 (3)**	0.7848 (4)	**0.7483 (2)**	**0.7490 (1)**
JesusFraile	0.7223 (4)	0.7827 (5)	0.7479 (3)	0.7478 (2)

Table 3: Snippet of Leaderboard sorted by Precision for SubTask-3

tasks.

In Subtask-A, the LGBM model on top of Tf-Idf performed the best for us with a F1-Score of 0.8684 while models like Naive Bayes, Logistic Regression, Random Forest and CatBoost on top of Tf-Idf were not too far away.

In Subtask-B, the CatBoost model on top of Tf-Idf performed the best with a score of 0.5604 while models like Naive Bayes, Logistic Regression and Random Forest were close with scores of 0.5482, 0.5577 and 0.5496 respectively.

In Subtask-C, the CatBoost model on top of Tf-Idf performed the best with a score of 0.7081, while models like Logistic Regression and Xgboost on top of Tf-Idf score 0.7075 and 0.6994 respectively and came very close.

We also performed fine-tuning using Transformers but the outcomes were inadequate, so we decided to use simpler models in order to achieve better performance.

We were eventually able to surpass the baseline F1 scores for Subtask-A: 0.708(BERT(Kenton and Toutanova, 2019)), Subtask-B: 0.554(BERT), and Subtask-C: 0.5495(Climate-BERT(Webersinke et al., 2021)) that were provided by the organizer.

Note that, all the scores mentioned are the performance on the hidden test set and directly taken from the system-run report provided on the competition website after finalized leaderboard 2, 3.

5 Conclusion

In summary, our research contributes crucial insights into hate speech and stance detection within climate activism. Employing classical machine learning models, such as LSTM, Xgboost, LGBM, and Catboost, revealed nuanced variations in performance. Notably, Catboost emerged as a strong performer, showcasing F1 scores of 0.5604 and 0.7081 for Subtask-B and Subtask-C. LGBM excelled in Subtask-A with an impressive F1 score of 0.8684. This study guides model selection for robust hate speech detection. As we conclude, our findings serve as a valuable resource for advancing tools aligned with the authentic dynamics of digital discourse in climate change activism.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

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*. ACM.

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. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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

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*.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14–25, Doha, Qatar. Association for Computational Linguistics.

Hristo Tanev, Nicolas Stefanovitch, Andrew Halterman, Onur Uca, Vanni Zavarella, Ali Hurriyetoglu, Bertrand De Longueville, and Leonida Della Rocca. 2023. Detecting and geocoding battle events from social media messages on the russo-Ukrainian war: Shared task 2, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 160–166, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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*.