Stance and Hate Event Detection in Tweets Related to Climate Activism - Shared Task at CASE 2024

Surendrabikram Thapa[1], **Kritesh Rauniyar**[2], **Farhan Ahmad Jafri**[3],
Shuvam Shiwakoti[2], **Hariram Veeramani**[4], **Raghav Jain**[5], **Guneet Singh Kohli**[6],
Ali Hürriyetoğlu[7], **Usman Naseem**[8]
[1]Virginia Tech, USA, [2]Delhi Technological University, India, [3]Jamia Millia Islamia, India,
[4]UCLA, USA, [5]University of Manchester, UK, [6]Thapar University, India,
[7]Wageningen Food Safety Research, Netherlands, [8]Macquarie University, Australia
[1]surendrabikram@vt.edu, [2]rauniyark11@gmail.com

Abstract

Social media plays a pivotal role in global discussions, including on climate change. The variety of opinions expressed range from supportive to oppositional, with some instances of hate speech. Recognizing the importance of understanding these varied perspectives, the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) at EACL 2024 hosted a shared task focused on detecting stances and hate speech in climate activism-related tweets. This task was divided into three subtasks: subtasks A and B concentrated on identifying hate speech and its targets, while subtask C focused on stance detection. Participants' performance was evaluated using the macro F1-score. With over 100 teams participating, the highest F1 scores achieved were **91.44%** in subtask C, **78.58%** in subtask B, and **74.83%** in subtask A. This paper details the methodologies of 24 teams that submitted their results to the competition's leaderboard.

1 Introduction

In an era dominated by digital communication, social media platforms serve as dynamic arenas where global conversations unfold in real-time. Twitter (now X)[1], in particular, with its diverse community, has emerged as a vital space for discussions on pressing global issues. Among these, the discourse surrounding climate change stands out as a critical topic that captivates the attention of users worldwide, with the masses expressing myriad opinions towards climate change (Fownes et al., 2018). As public awareness of climate change grows, and global movements like Friday For Future (FFF) (Wallis and Loy, 2021) that aim to draw policymakers' attention towards climate change house various social media platforms, the need for a nuanced understanding of the discourse around climate change in the digital realm becomes essential.

The escalating concern regarding climate change, coupled with the diverse range of discourse observed on Twitter, presents a distinctive amalgamation that encapsulates the intricate spectrum of emotions expressed by individuals toward this global issue. Within this spectrum lie various layers, including stance, which reflects individuals' inclinations toward specific viewpoints. As opinions are freely voiced, the prevalence of hate speech also emerges (Jafri et al., 2023; Thapa et al., 2023). Moreover, using humor in language is both an engaging and intricate mechanism for conveying ideas on pressing matters (Rauniyar et al., 2023). In order to unravel these complexities and enhance our understanding of online discussions concerning climate change, Shiwakoti et al. (2024) introduced a comprehensive multi-aspect dataset consisting of tweets related to climate change. This dataset includes five key aspects: the relevance of tweets to climate change, the stance conveyed in tweets, the presence of hate speech, the targets of such hate speech, and the presence of humor. Expanding upon this, we launched a shared task at the CASE 2024 workshop, held alongside EACL 2024, by utilizing this dataset. This shared task is subdivided into three subtasks: subtask A focuses on hate speech detection, subtask B revolves around identifying targets within hate speech, and subtask C delves into stance detection in tweets. Through this shared task, our objective is to foster active participation and cooperation in tackling the critical challenge of discerning stances on complex issues and identifying and curtailing hate speech within the digital sphere.

The subsequent sections of this paper are structured as follows: Section 2 presents an overview of the dataset used in our shared task. Section 3 out-

[1] In this paper, we have still used Twitter to refer X. The posts in X are referred to as tweets in our paper.

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 234–247
March 22, 2024 ©2024 Association for Computational Linguistics

lines the specific subtasks of the shared task. Furthermore, Section 4 explains about methodologies used by the teams submitting system description papers. Section 5 discusses a brief analysis of these system descriptions, while Section 6 serves as the concluding segment of the paper.

2 Dataset

In our shared task, we utilized the ClimaConvo dataset introduced by Shiwakoti et al. (2024). This dataset includes a total of 15,309 tweets centered around the climate crisis issue. The dataset has 6 major tasks, viz. Relevance, Stance, Hate Speech, Hate Direction, Hate Targets, and Humor. Only 10,407 of the tweets in this data were relevant, while the remaining 4,902 were non-relevant. We only used three tasks in our shared task: hate speech detection, hate targets detection, and stance detection. A total of 10,407 tweets were used for both subtask A and subtask C, while 999 tweets were used for subtask B in the shared task. For each subtask, we divided the dataset into stages for training, evaluating, and testing in a stratified way, keeping a proportionate split ratio of approximately 70-15-15. Table 1 represents the dataset statistics for the shared task.

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

Table 1: Dataset statistics for our shared task.

3 Shared Task Description

Hate speech refers to any form of communication that explicitly attacks an individual or a group based on their inherent characteristics, such as gender, religion, or race (Zhou et al., 2023). Stance describes the attitude or perspective expressed in a text towards a particular claim or topic (Hardalov et al., 2022; Rajaraman et al., 2023). Stance and hate detection can be used to analyze the structure of user interactions in conversational threads, providing valuable insights into the dynamics of online discussions.

3.1 Subtask A: Hate Speech Detection

This task involves determining whether a particular tweet exhibits hate speech. The dataset consists of tweets that have been annotated to indicate if the text includes hate speech or not. More precisely, the dataset is divided into two distinct classes: tweets that have been classified as ***Hate Speech*** and tweets that have been classified as ***No Hate Speech***.

3.2 Subtask B: Targets of Hate Speech Detection

This subtask aims to identify the target audience of hate speech within a specified set of hateful tweets. The subtask specifically focuses on classifying three specified targets outlined within the dataset, even though hate speech text may encompass different potential targets across multiple categories. The tweets in the dataset are labeled according to their targets, which can be classified as ***community***, ***individual***, or ***organization***. Therefore, we aim to identify these specific targets within tweets containing hate speech.

3.3 Subtask C: Stance Detection

The objective of the task is to identify various forms of stance within the specific tweet. This involves identifying three categories of stance in the dataset, labeled ‘Support’, ‘Oppose’, and ‘Neutral’.

4 Participants’ Methods

4.1 Overview

Out of the 100 participants who registered for the shared task, a total of 23 participants submitted scores for subtask A, 18 participants for subtask B, and 19 participants for subtask C. The leaderboards for these subtasks are provided in Table 2, Table 3, and Table 4. In subtask A, CUET_Binary_Hackers achieved the highest performance with an impressive F1-score of 91.44. Similarly, in subtask B, MasonPerplexity secured the top position with an F1-score of 78.58, while in subtask C, ARC-NLP emerged as the leader with the highest score of 74.83.

4.2 Methods

This section presents brief overviews of the system descriptions submitted by the participating teams in the shared task. These summaries are derived from the detailed approaches outlined in the participants’ system description papers.

4.2.1 Subtask A

CUET_Binary_Hackers (Farsi et al., 2024) proposed multiple numbers of machine learning (ML), deep learning (DL), transformers, and hybrid (combination of ML, DL, and LLM) based models with and without oversampling. Additionally, they used various feature extraction techniques, including Word2Vec (Pennington et al., 2014) and TF-IDF (Ramos et al., 2003; Adhikari et al., 2021) for machine learning and FastText (Joulin et al., 2016) and GloVe (Mikolov et al., 2013) for DL models. After incorporating the oversampling technique, they achieved best macro F1-score of 88% on SVM (Support Vector Machine) (Evgeniou and Pontil, 2001) and 88% on RF (Random Forest) (Louppe, 2015) machine learning models. However, without the oversampling technique, they achieved the best F1-score of 86% on SVM and 89% on RF model with TF-IDF and Word2Vec vectorizer respectively. In deep learning models with oversampling and by using Glove and FastText as vectorizers, BiGRU Cho et al. (2014) and CNN+BiGRU (Gehring et al., 2017) attained 80% and 90% F1-score respectively, but without oversampling with the same set of vectorizers, CNN+BiGRU achieved 91% (with GloVe) and 90% (with FastText) respectively. In transformer models with oversampling, mBERT (Devlin et al., 2019) and ClimateBERT (Webersinke et al., 2022) both achieved 91% F1-score and without oversampling, mBERT attained 91% F1-score. With this F1-score, the stood first on the leaderboard.

AAST-NLP (El-Sayed and Nasr, 2024a) used the **top-k** ensemble technique to achieve higher F1-score. Initially, they finetuned the bert variants, RoBERTa Liu et al. (2019), XLM-RoBERTa (Conneau et al., 2020) and HateBERT Caselli et al. (2021) on all of the datasets to attain the best results. They employed the 'Top-3' and 'Top-5' ensemble types, each of which used a different approach to attain the greatest F1-score. They obtain the maximum recall of 96.11% on HateBERT, the highest precision of 86.88% on RoBERTa, and the highest F1-score of 89.14% on **Top-5** ensemble approach. Of the 23 teams who participated in subtask A, their 'Top-5' ensemble approach, combining various BERT-based models, obtained the second position.

ARC-NLP (Kaya et al., 2024) used a combination of generative and encoder models, focusing on tweet-specific elements like hashtags, URLs, and emojis and employing optimization techniques like Optuna (Akiba et al., 2019). The work explores implementing three primary methods: the Encoder model, the Generative model, and the Hybrid model. The hybrid approach utilized a combination of the encoder model, such as BERTweet (Nguyen et al., 2020), and the generative model, such as Llama2 (Touvron et al., 2023). In subtask A, the hybrid model (BERTweet + Llama2) outperformed with an F1-score of 89.01% and secured third position in the leaderboard.

HAMiSoN Baselines (Montesinos and Rodrigo, 2024) evaluated the performance of the RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) models in the classification-based subtask and further investigated their performance when supplemented with external data. They combine two additional datasets such as the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019b) proposed at SemEval-2019 and Stance Detection Dataset (Mohammad et al., 2016b) released at the SemEval-2016 Task 6. They adopt preprocessing techniques such as replacing identifiers with special tokens and further decomposing hashtags into individual words to positively reinforce the learning process. They then analyze the performance of Hate speech classification in subtask A with RoBERTa and DeBERTa with the presence and absence of external datasets and report that standalone RoBERTa performed the best in subtask A with an F1-score of 88.86%, taking the fourth position in the leaderboard.

MasonPerplexity (Emran et al., 2024) used a weighted ensemble model combining the XLM-Roberta-Large (Conneau et al., 2020), HateBert (Caselli et al., 2021), and fBert (Sarkar et al., 2021), which were selected as the best three models from a pool of models tested. To handle the class imbalance challenge, the submission involved the concept of 'Back Translation', where text is translated from one language to another and then back to the original. This approach was explicitly applied to labels with lower representation in the training data, translating them through chains of multiple languages like *Xhosa to Twi to English, Lao to Pashto to Yoruba to English, Yoruba to Somali to Kinyarwanda to English and Zulu to Oromo to Shona to Tsonga to English*. This multi-language translation process introduces nuanced variations in

Rank	Team Name	Codalab Username	Accuracy	Recall	Precision	F1-score
1	CUET_Binary_Hackers (Farsi et al., 2024)	Asrarul_Hoque_Eusha	**96.35**	**91.73**	**91.16**	**91.44**
2	AAST-NLP (El-Sayed and Nasr, 2024b)	AhmedElSayed	95.71	86.54	92.31	89.14
3	ARC-NLP (Kaya et al., 2024)	kagankaya1	95.26	89.73	88.33	89.01
4	HAMiSoN-baselines (Montesinos and Rodrigo, 2024)	julioremo	95.39	87.97	89.81	88.86
5	MasonPerplexity (Gangul et al., 2024)	Sadiya_Puspo	95.52	86.89	91.12	88.85
6	HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024)	Raquel	95.33	86.55	90.53	88.40
7	Bryndza (Suppa et al., 2024)	mareksuppa	94.75	89.90	86.60	88.14
8	CUET_Binary_Hackers (Farsi et al., 2024)	SalmanFarsi	94.37	91.06	85.13	87.75
9	-	kojiro000	95.07	83.19	92.26	86.99
10	NLPDame (Christodoulou, 2024)	christiechris	94.62	84.32	89.09	86.49
11	CSI	RyszardStaurch	93.73	89.09	83.94	86.24
12	-	swatirajwal	94.43	84.21	88.33	86.11
13	-	refaat1731	94.94	79.68	96.07	85.56
14	-	d_rock	93.47	88.02	83.52	85.56
15	RACAI (Păiș, 2024)	pvf	94.37	82.79	89.07	85.55
16	Z-AGI Labs (Narayan and Biswal, 2024)	mrutyunjay_research	94.94	79.22	96.86	85.39
17	byteSizedLLM	mdp0999	94.17	80.16	90.44	84.29
18	JRC (Tanev, 2024)	htanev	94.05	77.79	92.46	83.10
19	-	Nikhil_7280	91.17	84.88	78.59	81.25
20	-	kriti7	88.92	91.18	75.66	80.26
21	Empty_heads	fayez94	87.96	50.00	43.98	46.80
21	pokemons	md_kashif_20	87.96	50.00	43.98	46.80
22	md_kashif_20	pakapro	50.38	52.28	50.97	42.42

Table 2: Sub-task A (Hate Speech Classification) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

the dataset, effectively enriching and diversifying the training examples for these underrepresented classes. Their approach helped them achieve a 5th Rank out of 22 submissions in subtask A, with an F1 score of 89%. They also tested with an ensemble of BERTweet-large, XLM-Roberta-Large, and fBERT which, yielded an F1-score of 84%.

HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024) took a Multi-task learning (MTL) approach with the help of multiple external datasets for classification problems across all 3 tasks. They took the hard parameter-sharing approach for MTL, using a different classification head for each task with a shared RoBERTa encoder. They performed extensive experimentation with multiple dataset combinations, and their best-performing model for hate speech detection (subtask A) achieved a F1-score of 88.40% and was ranked 6th among the 22 submissions. Although external datasets were used for experiments, their best performance was obtained using only the shared task dataset.

Bryndza (Suppa et al., 2024) investigates the utilization of GPT-4[2] (Brown et al., 2020), assessing its efficacy when used in both zero and few-shot learning (Hasan et al., 2023) and expanded through the incorporation of retrieval augmentation (Lewis et al., 2020) and re-ranking techniques (Mei et al., 2014). They discussed using the flashrank library (Damodaran, 2023) for re-ranking, aiming to enhance the model's performance in classification tasks. A suitable prompt was generated for subtask A by selecting a small sample of 30 Non-Hate and 30 Hate tweets and sending them to GPT-4. Chroma Vector Database[3] was utilized for retrieval augmentation to create an index of embeddings generated by pre-trained Sentence Transformer models[4], such as 'all-MiniLM-L6-v2' and 'all-mpnet-base-v2'. In subtask A, the model 'all-mpnet-base-v2' demonstrated notable effectiveness, yielding the ultimate submission with $k = 6$, k refers to the number of examples that can be chosen for retrieval augmentation, and the model achieved an F1-score of 88.14%.

NLPDame (Christodoulou, 2024) utilized parameter-efficient fine-tuning methodologies such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) and prompt tuning with the recently proposed Mistral 7B (Jiang et al., 2023) models for the evaluation of subtask A. They preprocess emojis to convert them into their equivalent textual representations, UTF-8 apostrophe encoding and normalization of identifiers such as user, URL, and Email and then adopt weighted cross entropy as the loss function. Further, the work compares the Mistral LLM's performance against the models previously proposed such as BERT, DistilBERT, RoBERTa, and ClimateBERT. The Mistral prompt

[2] `https://openai.com/gpt-4`

[3] `https://www.trychroma.com/`

[4] `https://www.sbert.net/docs/pretrained_models.html`

tuning approach achieved the highest F1-score of 86.4% in subtask A.

RACAI (Păiș, 2024) implemented a BERT-based model fused with hand-crafted features to detect hate speech (subtask A). They performed extensive pre-processing with the data which is superior to the dataset paper and significantly contributed to improving the performance of the model. The features included several raw hashtags, remaining hashtags after pre-processing, hashtags that were split during pre-processing, user mentions, URLs, and TF-IDF prediction. The final architecture is completed with the help of a Decision Tree (DT), which combines the LLM predictions with the features. However, their best-performing model as per the competition's evaluation metric (F1-score) turned out to be the plain fine-tuned BERT implementation which gave a F1-score of 85.55% and ranked 15 among the 22 submissions.

Z-AGI Labs (Narayan and Biswal, 2024) presented using conventional ML methods combined with contemporary DL techniques. In their study, the architecture of the DL model included a framework based on Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) with attention mechanisms (Vaswani et al., 2017). The Light Gradient Boosting Machine (LGBM) model (Ke et al., 2017), integrated with TF-IDF (Ramos et al., 2003) for feature extraction, yielded the most favorable results, achieving an F1-Score of 86.84%.

JRC (Tanev, 2024) participated in Sub-task A: 'Hate Speech Detection', where they employed a pure lexicon-based method (Gitari et al., 2015), avoiding statistical classifiers, and achieved moderate performance compared to other participants. Their model ranked 18 out of 22 participants, with an F1-score of 83.10%.

4.2.2 Subtask B

MasonPerplexity (Emran et al., 2024) in their approach for subtask B, they used a distinct set of individual models comprising XLM-R (Conneau et al., 2020), BERT-Base (Devlin et al., 2019), and BERTweet-Large (Nguyen et al., 2020). Among these, BERTweet-Large was particularly notable, achieving outstanding results with an accuracy of 91.33%, precision of 81.33%, and recall of 78.23%. This performance led to a test F1-score of 79%, securing the 1st rank among 18 submissions for the task. The research team also implemented the 'Back Translation' technique to address class imbalance in the dataset. This involved translating texts from underrepresented labels through various language sequences, such as *Xhosa to Twi to English, Lao to Pashto to Yoruba to English, Yoruba to Somali to Kinyarwanda to English, and Zulu to Oromo to Shona to Tsonga to English*, and then back to English. Introducing nuanced linguistic variations significantly enriched and diversified the training data, effectively improving the model's ability to handle underrepresented classes.

Bryndza (Suppa et al., 2024) presented the performance of different models with GPT-4 used for detecting the targets of hate speech in tweets. Within subtask B, the retrieval-augmentation approach utilizing hte 'all-MiniLM-L6-v2' model where $k = 6$, produced the most favorable outcomes, achieving an F1-score of 77.61% and second position in the leaderboard.

AAST-NLP (El-Sayed and Nasr, 2024a) presented the **top-k** ensemble strategy to reach higher F1-score. To get the best results, they first tweaked the BERT types on all datasets: RoBERTa, XLM-RoBERTa, and HateBERT. In this task, they also experimented with two named entity recognition (NER) modules: SpaCy [5] and BERT-based NER. While ORG and NoORG landmarks were extracted using SpaCy, names were extracted more effectively by the BERT-based NER. They employed the 'Top-3' and 'Top-5' ensemble styles, each taking a distinct method to get the highest F1-score. With the **Top-5** ensemble strategy, they achieve the maximum of all three metrics: F1-score of 76.65%, recall of 76.89%, and accuracy of 77.06%. Their 'Top-5' ensemble technique, which integrates multiple BERT-based models, secured a second place among the eighteen teams who took part in subtask B.

ARC-NLP (Kaya et al., 2024) focused on BERTweet, which is an encoder-based technique for classification, and achieved the highest F1-score of 76.38% among the other four models. Another model, BERTweet+NER (Nguyen et al., 2020; Ozcelik and Toraman, 2022), is a hybrid formed by combining encoder and generative models, and it also scored an F1-score of 75.00%.

CUET_Binary_Hackers (Farsi et al., 2024) presented various models and feature extraction tech-

[5] https://spacy.io/

Rank	Team Name	Codalab Username	Accuracy	Precision	Recall	F1-score
1	MasonPerplexity (Gangul et al., 2024)	Sadiya_Puspo	**91.33**	**81.33**	**78.23**	**78.58**
2	Bryndza (Suppa et al., 2024)	mareksuppa	92.67	78.13	77.61	77.61
3	AAST-NLP (El-Sayed and Nasr, 2024b)	AhmedElSayed	91.33	76.89	77.06	76.65
4	ARC-NLP (Kaya et al., 2024)	kagankaya1	91.33	77.28	75.88	76.38
5	-	kojiro000	91.33	73.23	77.06	74.88
6	CUET_Binary_Hackers (Farsi et al., 2024)	SalmanFarsi	90.00	74.31	75.33	74.33
7	-	amr8ta	90.00	71.29	78.26	73.65
8	HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024)	Raquel	90.00	71.54	75.33	73.29
9	-	swatirajwal	89.33	67.39	69.78	68.48
10	HAMiSoN-baselines (Montesinos and Rodrigo, 2024)	julioremo	87.33	64.71	73.64	65.88
11	NLPDame (Christodoulou, 2024)	christiechris	84.00	61.51	72.85	61.06
12	byteSizedLLM	mdp0999	88.67	52.33	62.46	55.80
13	-	Nikhil_7280	88.00	51.66	61.01	54.96
14	EmptyMind	empty_box	87.33	52.39	56.04	54.07
15	Z-AGI Labs (Narayan and Biswal, 2024)	mrutyunjay_research	86.00	50.71	51.97	51.33
16	Team +1	pakapro	30.00	33.53	38.80	24.58
17	-	kriti7	7.33	13.95	4.91	7.18
18	pokemons	md_kashif_20	0.00	0.00	0.00	0.00

Table 3: Sub-task B (Targets of Hate Speech Detection) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

niques similar to their approach in subtask A. The best F1-score of 74% were obtained with the oversampling technique with mBERT and DistillBERT models.

HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024) leveraged MTL for all 3 tasks with the hard parameter-sharing approach, using a different classification head for each task and a shared RoBERTa encoder for all. They also performed extensive experimentation with external datasets, and their best-performing model for hate target detection achieved a F1-score of 73.29% and ranked 8th among 18 submissions. This score was obtained using the shared task dataset and the target identification task dataset from OLID (Zampieri et al., 2019a), which is an extensive offensive language detection dataset.

HAMiSoN Baselines (Montesinos and Rodrigo, 2024) Similar to subtask A, they analyze the performance of the RoBERTa and DeBERTa models in the three classification-based subtasks with external data augmentation using the two additional datasets such as Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019b) proposed at SemEval-2019 and Stance Detection Dataset (Mohammad et al., 2016b) released at the SemEval-2016 Task 6. They continue to reuse the preprocessing techniques, such as replacing identifiers with special tokens and hashtag decomposition into simple words to improve the downstream model's prediction. Based on their performance report of Target Identification subtask B, they note that the standalone RoBERTa with external data performed the best in subtask B with an F1-score of 70.17%.

NLPDame (Christodoulou, 2024) Similar to their approach in subtask A, they adopted LoRA and prompt tuning methods based on Mistral for the target identification subtask B. They reuse the pre-processing techniques such as emoji conversion to their equivalent textual representations, apostrophe encoding in UTF-8 style, and normalization of identifiers of identifiers. They adopted weighted cross entropy as the loss function for the three-class classification task with inherent task imbalance. Finally, they discuss Mistral LLM's performance compared to transformer models like BERT, DistilBERT, RoBERTa, and ClimateBERT. The prompt tuning approach with Mistral yielded them the highest F1-score of 61.0% in subtask B.

Z-AGI Labs (Narayan and Biswal, 2024) worked on various ML and DL approaches where they used TF-IDF for the feature extraction. The CatBoost (Prokhorenkova et al., 2018) model exhibited superior performance, achieving an F1-score of 56.04%. In comparison, models such as Naive Bayes, LR, and RF closely followed with F1-scores of 54.82%, 55.77%, and 54.95%, respectively.

4.2.3 Subtask C

ARC-NLP (Kaya et al., 2024) used the optimized version of the BERTweet model. This model outperformed other encoder models in stance detection; it employed a short input tokenization length (96 tokens) and incorporated special tokens for tweet-specific elements. The highest macro F1-score was achieved by the BERTweet model,

Rank	Team Name	Codalab Username	Accuracy	Precision	Recall	F1-score
1	ARC-NLP (Kaya et al., 2024)	kagankaya1	**74.90**	**78.48**	**72.26**	**74.83**
2	HAMiSoN-Generative (Fraile-Hernandez and Peñas, 2024)	JesusFraile	74.78	78.27	72.23	74.79
3	IUST (Mahmoudi and Eetemadi, 2024)	gh_mhdi	73.11	78.63	71.45	74.47
4	HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024)	Raquel	74.33	77.02	72.42	74.02
5	AAST-NLP (El-Sayed and Nasr, 2024b)	AhmedElSayed	74.39	79.31	70.78	73.98
6	MasonPerplexity (Gangul et al., 2024)	Sadiya_Puspo	73.69	77.80	70.90	73.73
7	-	kojiro000	73.43	77.44	70.89	73.58
8	-	refaat1731	72.22	77.49	70.06	73.15
9	HAMiSoN-baselines (Montesinos and Rodrigo, 2024)	julioremo	74.01	78.17	70.36	73.13
10	-	Nikhil_7280	71.90	76.62	68.13	70.81
11	-	swatirajwal	67.86	70.83	70.05	70.26
12	Bryndza (Suppa et al., 2024)	mareksuppa	71.19	68.72	71.23	69.33
13	NLPDame (Christodoulou, 2024)	christiechris	66.52	71.16	67.94	69.30
14	byteSizedLLM	mdp0999	65.24	72.55	66.85	69.10
15	CUET_Binary_Hackers (Farsi et al., 2024)	SalmanFarsi	66.13	69.08	66.91	67.94
16	Z-AGI Labs (Narayan and Biswal, 2024)	mrutyunjay_research	69.08	79.26	62.94	63.72
17	Team +1	pakapro	32.71	32.66	31.51	28.98
18	-	ankitha11	0.38	1.32	0.16	0.29
19	pokemons	md_kashif_20	0.00	0.00	0.00	0.00

Table 4: Sub-task C (Stance Detection) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

which scored 74.83%. This score is slightly higher than the other models tested for the same subtask, with DeBERTa (He et al., 2021) coming close with an F1-score of 73.85%. The optimization of the BERTweet model, focusing on tweet-specific elements, was key to its top performance. For subtask C, BERTweet outperformed all the other models in the leaderboard securing the first position.

HAMiSoN-Generative (Fraile-Hernandez and Peñas, 2024) implemented variants/modifications of the Llama 2 7B generative LLM for stance prediction (subtask C). 3 of the 4 variants of the Llama 2 7B used are out-of-the-box chatbot models, but by using specific input formats, these models were adapted to be used in classification tasks. They also used an external data source (Mohammad et al., 2016a), which is related to the stance detection, to train and boost their models' performance. Despite the models used being chatbot models, they were able to achieve 2nd position in the stance detection sub-task among 19 submissions with an impressive F1-score of 74.79%.

IUST (Mahmoudi and Eetemadi, 2024) evaluated models such as BERT, RoBERTa, BERTWeet, XLM-RoBERTa, and DEBERTA for the three subtasks. Data augmentation strategies such as synonym substitution and Round-trip translation and German as the back translation language using nlpaug library[6] were adopted as part of the pipeline. The main focus of this work is to focus on optimal hyperparameter selection from the search space definition comprising of the optimizers, loss functions (Focal loss/Weighted cross-entropy loss), cleaning strategies, classification layer choices of a Fully Connected Layer/ Convolutional Neural Network (CNN) head architectures were investigated while demonstrating that CNN classifier heads performed across all their cleaning strategy/Embedding model based pipelines. The cleaning strategy of removing URL and username identifiers, in addition to stochastic gradient descent optimizer and CNN classifier head, was demonstrated to have achieved the highest F1-scores of 73.97%, 74.47% based on XLM-ROBERTa and BERTweet based systems on the Climate stance detection task.

HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024) used a RoBERTa-based Multi-task learning approach for all 3 tasks by a RoBERTa shared encoder for all and a different classification head for each task. They used external datasets and performed multiple experiments with various dataset combinations. Combining the shared task dataset and the OLID (Zampieri et al., 2019a), they achieved their best performance with a F1-score of 74.02% and ranked 4th among 19 submissions in the stance detection task.

AAST-NLP (El-Sayed and Nasr, 2024a) leverage the **top-k** ensemble strategy to reach higher F1-score. To get the best results, they first tweaked the bert types on all datasets: RoBERTa, XLM-RoBERTa, and HateBERT. In this subtask, they only employed 'Top-5' ensemble styles, which made them get the highest F1-score. In this subtask, their RoBERTa model achieves the best precision value of 71.69% and with the use **Top-5** ensem-

[6] https://github.com/makcedward/nlpaug

ble strategy, they achieve the maximum recall and f1-score metrics value of 79.31% and 73.98% respectively. In subTask-C, they attained the fifth position on 18 participating teams by using the 'Top-5' ensemble technique.

MasonPerplexity (Emran et al., 2024) implemented a variety of models including BERTweet-large, BERT base, and BERTweet base. Of these, the BERTweet base model stood out, achieving the highest F1-score. Their system ranked 6th out of 19 submissions. The performance of the different models experimented with are as follows: GPT3.5 Zero Shot prompting had a Test F1-score of 63%, GPT-3.5 Few Shot prompting achieved a Test F1-score of 67%, BERT- BASE scored a Test F1-score of 69%, BERTweet-LARGE attained a Test F1-score of 70%, and BERTweet-Base led the group with a Test F1-score of 74%.

HAMiSoN Baselines (Montesinos and Rodrigo, 2024) Similar to subtask A and subtask B, they report the performance of the RoBERTa and DeBERTa models with external data augmentation using the two additional datasets such as OLID ((Zampieri et al., 2019b)) proposed at SemEval-2019 and Stance Detection Dataset ((Mohammad et al., 2016b)) released at the SemEval-2016 Task 6 reusing the similar pre-processing techniques they note that the standalone RoBERTa without external data performed the best in subtask C with a F1-score of 74.95%.

Bryndza (Suppa et al., 2024) made the use of 'all-mpnet-base-v2' model with GPT-4 API, which was highly effective, leading to its selection for the final submission. With $k = 8$, it achieved an F1-score of 69.33%, demonstrating its strong performance in classifying stance.

NLPDame (Christodoulou, 2024) Similar to their approach in subtask A and subtask B, they adopted LoRA and prompt tuning Parameter efficient fine-tuning methods based on Mistral and reused the pre-processing techniques such as emojis conversion to their equivalent textual representations, apostrophe encoding in UTF-8 style and normalization of identifiers of key identifiers part of the samples like user, URL, Email for the target identification subtask B. Following this approach, they conclude that superior performance of Mistral LLMs continues to emerge again in subtask C, similar to the previous subtasks as compared to the transformer models like BERT, DistilBERT, RoBERTa, and ClimateBERT fetching them the highest F1-score of 69.3% in subtask C.

CUET_Binary_Hackers (Farsi et al., 2024) presented various learning models with diverse feature engineering and oversampling techniques. The best results were obtained with oversampling techniques. DistillmBERT, ClimateBERT and BiGRU (with Glove embeddings) gave a same F1-score of 67%. F1-scores of 31% without oversampling and 62% with oversampling were obtained using their hybrid model. mBERT+BiLSTM+CNN (Mustavi Maheen et al., 2022).

Z-AGI Labs (Narayan and Biswal, 2024) focused on stance detection, the CatBoost model based on TF-IDF was the top performer. This model achieved the highest F1-score of 70.80%, indicating its effectiveness in accurately categorizing stances in the context of climate activism. The paper explores the strengths of the model and its proficiency in handling the complexities of stance detection, compared to other models like Logistic Regression (Indra et al., 2016) and XGBoost (Haumahu et al., 2021), which also showed close performance.

HAMiSoN-Ensemble (Rodriguez-Garcia et al., 2024) present an ensemble approach of Roberta, a generative LLM - Llama 2, and Multi-task learning for stance detection. For the Llama 2, they used the Llama-2 7B Chat model with necessary modifications to adapt it to classification tasks. As for Multi-task learning, they used a RoBERTa-based model. They also used external data to improve their model performance but do not seem to show an added advantage over just using the competition dataset. Although they performed a majority voting ensemble approach of the 3 models, their best-performing model was the fine-tuned Roberta model, which achieved a F1-score of 73.13%. However, as mentioned in their paper, on a post-competition analysis, through some modifications in their approach, their ensemble system achieved a F1-score of 75.29%, which surpasses their RoBERTa-based system.

5 Discussion

The results and methodologies presented by the teams participating in this shared task offer valuable insights into the current state-of-the-art in hate speech detection, target identification, and stance detection. These tasks are essential in understanding the dynamics of online discourse, particularly on social media platforms. A notable trend across all subtasks is the heavy reliance on transformer-based models, particularly BERT and its variants. These models have shown exceptional capability in understanding the intricacies of natural language, especially in informal and idiosyncratic texts commonly found on social media. Their success underlines the importance of advanced models in handling the complexities of language in these contexts. Ensemble and hybrid approaches have also been prevalent, adopted by teams like AAST-NLP (El-Sayed and Nasr, 2024a) and CUET_Binary_Hackers (Farsi et al., 2024). Another critical aspect highlighted by several teams is handling class imbalance in datasets. The use of external datasets to enrich training data, as seen in the approaches of HAMiSoN-MTL (Rodriguez-Garcia and Centeno, 2024) and HAMiSoN-Generative (Fraile-Hernandez and Peñas, 2024), indicates a growing recognition of the value of diverse and expansive data sources. This approach can lead to better generalization and robustness of the models. Preprocessing and feature engineering also play a crucial role, as demonstrated by teams like MasonPerplexity (Emran et al., 2024), and Bryndza (Suppa et al., 2024). The way data is prepared and presented to models can significantly impact their effectiveness, highlighting the importance of meticulous data handling. Incorporating the latest advancements in LLMs further enriches the discussion of shared tasks' outcomes and future directions. The use of LLMs, as demonstrated by teams like Bryndza (Suppa et al., 2024) and MasonPerplexity (Emran et al., 2024), marks a significant shift in the approach to understanding and processing natural language on social media platforms. Despite these advances, several challenges and potential future directions emerge. Ensuring that models perform well across different contexts remains a significant challenge, given the variability in expressions of hate speech and stances. Additionally, the subtlety and ambiguity in language use, especially in these domains, continue to pose significant hurdles.

6 Conclusion

The shared task at the CASE 2024 workshop has made significant strides in advancing our understanding of hate speech detection, target identification, and stance detection in social media contexts, focusing on Twitter conversations about climate change. The diversity of approaches employed by the participants, predominantly centered around sophisticated transformer-based models like BERT and its variants, demonstrates the complexity of analyzing online discourse. However, this field of study still faces significant challenges, including ensuring the adaptability of models across various contexts, refining language processing to capture subtle nuances, and navigating the ethical implications of automated content analysis. This task has provided a comprehensive benchmark for current methodologies and set the stage for future research in the rapidly evolving domain of NLP, emphasizing the need for continued innovation in understanding the complexities of digital communication.

Acknowledgements

We would like to acknowledge the support of numerous reviewers who helped to provide reviews for this shared task. We would also like to acknowledge the support of various researchers who helped us to advertise this shared task. Moreover, this work is supported by the European Research Council Politus Project (ID:101082050) and European Union's HORIZON projects EFRA (ID: 101093026) and ECO-Ready (ID: 101084201).

Broader Impact

The broader impact of the CASE 2024 workshop's shared task extends across various domains, significantly influencing social media moderation, public policy, academic research, ethical AI development, and more. This research aids in enhancing content moderation on social media platforms, helping to create safer and more inclusive online communities by effectively identifying and mitigating harmful content. In public policy and awareness, insights from stance detection, particularly on critical issues like climate change, are invaluable for policymakers and advocacy groups, aiding in developing resonant communication strategies and informed policies. The task fosters interdisciplinary collaboration, merging expertise from linguistics, computer science, sociology, and environmental studies, enriching academic research and encouraging innovative approaches in NLP and social media analysis. It also contributes to the broader discourse on ethical AI, emphasizing the need for transparent and accountable AI systems, especially in sensitive areas like hate speech analysis. The showcasing of advanced models like GPT-4 and BERT highlights the continual evolution of NLP technologies, opening doors for more sophisticated and context-aware AI tools. Given the global nature of social media, the advancements in NLP and AI have the potential to impact digital communication worldwide. This shared task contributes to possible scalable solutions that can be adapted across different languages and cultures.

References

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Christina Christodoulou. 2024. NLPDame at Climate-Activism 2024: Mistral Sequence Classification with PEFT for Hate Speech, Targets and Stance Event Detection. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

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.

Prithiviraj Damodaran. 2023. Flashrank, lightest and fastest 2nd stage reranker for search pipelines. https://doi.org/10.5281/zenodo.10426927.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ahmed El-Sayed and Omar Nasr. 2024a. AAST-NLP at ClimateActivism 2024: Ensemble-Based Climate Activism Stance and Hate Speech Detection : Leveraging Pretrained Language Models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Ahmed El-Sayed and Omar Nasr. 2024b. AAST-NLP at Multimodal Hate Speech Event Detection 2024 : A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Al Nahian Bin Emran, Amrita Ganguly, Sadiya Sayara Chowdhury Puspo, Dhiman Goswami, and Md Nishat Raihan. 2024. MasonPerplexity at ClimateActivism 2024: Integrating Advanced Ensemble Techniques and Data Augmentation for Climate Activism Stance and Hate Event Identification. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Theodoros Evgeniou and Massimiliano Pontil. 2001. Support vector machines: Theory and applications. volume 2049, pages 249–257.

Salman Farsi, Asrarul Hoque Eusha, and Mohammad Shamsul Arefin. 2024. CUET_Binary_Hackers at ClimateActivism 2024: A Comprehensive Evaluation and Superior Performance of Transformer Models in Hate Speech Detection and Stance Classification for Climate Activism. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Jennifer R Fownes, Chao Yu, and Drew B Margolin. 2018. Twitter and climate change. *Sociology Compass*, 12(6):e12587.

Jesus M. Fraile-Hernandez and Anselmo Peñas. 2024. HAMiSoN-Generative at ClimateActivism 2024: Stance Detection using generative large language models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Amrita Gangul, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, Dhiman Goswami, and Marcos Zampieri. 2024. Mason-Perplexity at Multimodal Hate Speech Event Detection 2024: Hate Speech and Target Detection Using Transformer Ensembles. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. #metooma: Multi-aspect annotations of tweets related to the metoo movement. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):209–216.

Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. 2017. A convolutional encoder model for neural machine translation.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783*.

JP Haumahu, SDH Permana, and Y Yaddarabullah. 2021. Fake news classification for indonesian news using extreme gradient boosting (xgboost). In *IOP Conference Series: Materials Science and Engineering*, volume 1098, page 052081. IOP Publishing.

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*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

ST Indra, Liza Wikarsa, and Rinaldo Turang. 2016. Using logistic regression method to classify tweets into the selected topics. In *2016 international conference on advanced computer science and information systems (icacsis)*, pages 385–390. IEEE.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines.

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*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Ahmet Kaya, Oguzhan Ozcelik, and Cagri Toraman. 2024. ARC-NLP at ClimateActivism 2024: Stance and Hate Speech Detection by Generative and Encoder Models Optimized with Tweet-Specific Elements. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. 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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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*.

Gilles Louppe. 2015. Understanding random forests: From theory to practice.

Ghazaleh Mahmoudi and Sauleh Eetemadi. 2024. IUST at ClimateActivism 2024: Towards Optimal Stance Detection: A Systematic Study of Architectural Choices and Data Cleaning Techniques. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Tao Mei, Yong Rui, Shipeng Li, and Qi Tian. 2014. Multimedia search reranking: A literature survey. *ACM Computing Surveys (CSUR)*, 46(3):1–38.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Julio Reyes Montesinos and Alvaro Rodrigo. 2024. HAMiSoN-baselines at ClimateActivism 2024: A Study on the Use of External Data for Hate Speech and Stance Detection. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Syed Mustavi Maheen, Moshiur Rahman Faisal, Md. Rafakat Rahman, and Md. Shahriar Karim. 2022. Alternative non-BERT model choices for the textual classification in low-resource languages and environments. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 192–202, Hybrid. Association for Computational Linguistics.

Nikhil Narayan and Mrutyunjay Biswal. 2024. Z-AGI Labs at ClimateActivism 2024: Stance and Hate Event Detection using Tf-Idf and LSTM. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

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.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

Vasile Păiș. 2024. RACAI at ClimateActivism 2024: Improving Detection of Hate Speech by Extending LLM Predictions with Handcrafted Features. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Kanagasabai Rajaraman, Hariram Veeramani, Saravanan Rajamanickam, Adam Maciej Westerski, and Jung Jae Kim. 2023. Semantists at imagearg-2023: Exploring cross-modal contrastive and ensemble models for multimodal stance and persuasiveness classification. In *Proceedings of the 10th Workshop on Argument Mining*, pages 181–186.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*.

Raquel Rodriguez-Garcia and Roberto Centeno. 2024. HAMiSoN-MTL at ClimateActivism 2024: Detection of Hate Speech, Targets and Stance using Multitask Learning. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Raquel Rodriguez-Garcia, Julio Reyes Montesinos, Jesus M. Fraile-Hernandez, and Anselmo Peñas. 2024. HAMiSoN-Ensemble at ClimateActivism 2024: Ensemble of RoBERTa, Llama 2 and Multitask for Stance Detection. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Annika Stechemesser, Anders Levermann, and Leonie Wenz. 2022. Temperature impacts on hate speech online: evidence from 4 billion geolocated tweets from the usa. *The Lancet Planetary Health*, 6(9):e714–e725.

Marek Suppa, Daniel Skala, Daniela Jass, Samuel Sucik, and Andrej Svec. 2024. Bryndza at ClimateActivism 2024: Stance, Target and Hate Event Detection via Retrieval-Augmented GPT-4. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Hristo Tanev. 2024. JRC at ClimateActivism 2024: Lexicon-based Detection of Hate Speech. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

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.

Hannah Wallis and Laura S Loy. 2021. What drives pro-environmental activism of young people? a survey study on the fridays for future movement. *Journal of Environmental Psychology*, 74:101581.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2023. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, 29(5):1247–1274.

A Related Works

In a range of social contexts, a link has been shown between weather and offline abuse. Concurrently, there is a significant number of online social issues as a result of almost every element of daily life becoming rapidly digitalized. Hate speech on the internet has become a major issue and has been demonstrated to exacerbate mental health issues, particularly in youth and marginalized communities (Stechemesser et al., 2022). ALDayel and Magdy (2021) explore the new trends and diverse uses of stance detection on social media. Stance detection on social media is a developing opinion-mining paradigm for various political as well as social purposes in which sentiment analysis may not be the best approach. Zampieri et al. (2019a) gathered the Offensive Language Identification Dataset (OLID), a new dataset containing tweets annotated for offensive content using a fine-grained three-layer annotation scheme, and compared the effectiveness of various machine learning models on OLID. They target a variety of different types of offensive content. Gautam et al. (2020) presented a dataset of 9,973 tweets on the MeToo movement that were manually annotated for five different language dimensions: dialogue acts, sarcasm, hate speech, relevance, and stance. The data was then examined in terms of keywords, label correlations, and geographical distribution. Mollas et al. (2022) provided access to 'ETHOS' (multi-labEl haTe speecH detectiOn dataSet), a textual dataset consisting of two variants: binary and multi-label, based on comments from Reddit and YouTube that were verified by the Figure-Eight crowdsourcing platform. Additionally, the annotation protocol—an active sampling process—that was utilized to create this dataset—was presented, in addition.

B Evaluation and Competition

This section describes the structure of our competition, along with the methodology used to determine ranks and other relevant data.

B.1 Evaluation Metrics

To evaluate the effectiveness of the participants' contributions, we used macro F1-score, accuracy, precision, and recall. The participants' ranks were determined using the macro F1-score sorting approach.

B.2 Competition Setup

We used the Codalab[7] to organize our competition. The competition consisted of two phases: an assessment phase where competitors got comfortable with the Codalab system and a testing phase where performance was used to determine the final ranking on the scoreboard.

Registration: A total of 100 participants registered for our competition, and the diverse array of email domains used indicated its success in attracting individuals from various parts of the world. Among the registrants, 23 teams submitted their predicted outcomes, reflecting active engagement and interest in the competition.

Competition Timelines: On November 1, 2023, training and evaluation data were made available, marking the commencement of the competition. The first half was evaluation-focused, with the main goal being to familiarize participants with Codalab. Participants were given access to the labels of the evaluation information in order to help with this process. The test phase then began on November 30, 2023, when test data was provided without any ground truth labels. The test session, which was originally scheduled to finish on January 5, 2024, was extended until January 7, 2024, in response to requests from many participants, displaying flexibility in meeting participant demands. In addition, it was finally determined that system description papers must be submitted by January 13, 2024. Participants were given a certain period to provide their system designs and approaches by this crucial deadline. The well-planned schedule made it possible for the competition to go through its phases thoroughly and organized, giving participants plenty of time to become involved, get familiar with one another, and submit their thoughtful submissions by the deadlines.

[7] The competition page can be found here: https://codalab.lisn.upsaclay.fr/competitions/16206.