CUET_Binary_Hackers at ClimateActivism 2024: A Comprehensive Evaluation and Superior Performance of Transformer-based Models in Hate Speech Event Detection and Stance Classification for Climate Activism

Salman Farsi, Asrarul Hoque Eusha and Mohammad Shamsul Arefin
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{salman.cuet.cse, asrar2860}@gmail.com, sarefin@cuet.ac.bd

Abstract

The escalating impact of climate change on our environment and lives has spurred a global surge in climate change activism. However, the misuse of social media platforms like Twitter has opened the door to the spread of hatred against activism, targeting individuals, organizations, or entire communities. Also, the identification of the stance in a tweet holds paramount significance, especially in the context of understanding the success of activism. So, to address the challenge of detecting such hate tweets, identifying their targets, and classifying stances from tweets, this shared task introduced three sub-tasks, each aiming to address exactly one mentioned issue. We participated in all three sub-tasks and in this paper, we showed a comparative analysis between the different machine learning (ML), deep learning (DL), hybrid, and transformer-based models. Our approach involved proper hyper-parameter tuning of models and effectively handling class imbalance datasets through data oversampling. Notably, our fine-tuned m-BERT achieved a macro-average $f1$ score of 0.91 in sub-task A (Hate Speech Detection) and 0.74 in sub-task B (Target Identification). On the other hand, Climate-BERT achieved a $f1$ score of 0.67 in sub-task C. These scores positioned us at the forefront, securing 1^{st}, 6^{th}, and 15^{th} ranks in the respective sub-tasks. The detailed implementation information for the tasks is available in the GitHub [1].

1 Introduction

Over the decades, climate change has evolved into a pressing issue for nature and all Earth's species, with alarming consequences. Reports from the Intergovernmental Panel on Climate Change (IPCC) confirm that climate change is resulting in more frequent and severe weather events, including heatwaves, droughts, and floods [2]. These events can lead to crop failures, food shortages, displacement of people, melting of glaciers and ice caps, rising sea levels, and increased coastal flooding.

Preserving a harmonious climate is vital for ensuring balanced ecosystems, optimal temperature conditions, and biodiversity (Weiskopf et al., 2020; Mikhaylov et al., 2020). This urgent issue has spurred people worldwide to voice their concerns and participate in a growing number of climate change activism events on a global scale (Damoah et al., 2023). These events aim to raise awareness about the impact of climate change and the urgent need for action. One such prominent movement is 'FridayForFuture' (FFF), initiated by Greta Thunberg, a Swedish schoolgirl, in August 2018, to exert pressure on policymakers to take necessary actions against climate change (Spaiser et al., 2022; Neas et al., 2022). Other notable climate activism movements, including 'Extinction Rebellion', 'Earth Strike', and 'Climate Justice Now', have further fueled the global movement against climate change (Gunningham, 2019; Schlosberg and Collins, 2014; Laux, 2021).

However, contemporary activism extends beyond street protests to online platforms, with social media users expressing their thoughts on climate movements through tweets and comments. But some people share hateful, aggressive, and humorous tweets targeting activism (Thapa et al., 2024). Hate speech not only undermines the objectives of activism but also poses a threat to the well-being of individuals, organizations, and communities involved in the movement (Arce-García et al., 2023). Whereas stance detection in text is also a vital component in assessing the dynamics of protests and activism. It helps understand whether activist movements and protests are being supported or opposed (Shiwakoti et al., 2024). Despite numerous studies conducted in recent years on identifying hate speech and its targets in social media text, this context in climate activism remains an under-explored

[1] https://github.com/Salman1804102/CASE-EACL-2024
[2] https://www.ipcc.ch/report/ar6/wg1/chapter/chapter-11/

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 145–155

domain (Parihar, Anil Singh and Thapa, Surendrabikram and Mishra, Sushruti, 2021; MacAvaney et al., 2019; Kovács et al., 2021). As these events serve as crucial platforms for promoting environmental awareness and policy changes, there is a need for a comprehensive understanding of the stance and mitigation strategy for hateful tweets. As contributors to this endeavor, our principal contributions are delineated below:

- We introduced and advocated for the utilization of BERT models by effectively handling the class imbalance data, leveraging their capabilities to classify textual content.
- By delving into diverse methodologies, we seek to provide valuable insights that can inform the development of more robust systems for addressing the intricacies of climate activism events on social media platforms.

The later part of the paper is organized as follows: Section 3 provides the task and dataset description, Section 4 outlines the methodology, Section 5 presents the result analysis, and Section 6 delves into error analysis for each task. Lastly, Section 7 encapsulates the conclusion.

2 Related Work

2.1 Hate Speech Detection

Over time, numerous research efforts have been dedicated to the detection and classification of hate speech, employing various methodologies. In an earlier study (Malmasi and Zampieri, 2017), a machine-learning approach was adopted, utilizing an SVM classifier with lexical features on a dataset comprising 14,509 English tweets. The results indicated a 78% accuracy using the 4-gram model. The exploration of machine learning methods continued in another study (Davidson et al., 2017), where Logistic Regression (LR) outperformed Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF) in identifying hate speech within a Twitter-based hate speech datasets.

As the popularity of deep learning algorithms grew, Zhang and Luo (2019) aimed to enhance the semantic understanding of hate speech. They introduced a CNN+(skipped-CNN) model, which showcased better performance compared to the CNN+GRU model across various publicly available Twitter datasets. Another deep learning-based study (Badjatiya et al., 2017) combined embeddings learned from LSTM with gradient-boosting decision trees, which achieved a higher $f1$ score of 93% in hate speech detection. The study also involved a comparative analysis utilizing various feature extraction methods such as character n-grams, word n-grams, fastText, GloVe, and Bag-of-words for LR, DT, and SVM. However, with the advent of transformer-based models like BERT, research trends shifted towards leveraging these models due to their capability to capture intricate semantic meanings in textual context. Mozafari et al. (2020) proposed BERT+LSTM, BERT+CNN, and BERT+Nonlinear-layers models for hate speech detection. Their BERT+CNN architecture demonstrated $f1$ scores of 88% and 92% for the Waseem (Waseem and Hovy, 2016) and Davidson (Davidson et al., 2019) hate datasets.

2.2 Hate Speech Target Identification

In the realm of hate speech target classification, researchers have extended their focus beyond merely detecting hate speech to the classification of hate speech targets. The study (Kurniawan and Budi, 2020) employed a labeled dataset of hate tweets in Indonesia, distinguishing between individual and group-targeted hate. Their work utilized word n-grams, Bag-of-words, and TF-IDF for machine learning models. Ultimately, the findings revealed that SVM surpassed NB and RF, achieving an impressive $f1$ score of 0.84772 with TF-IDF. In another work, (Shvets et al., 2021) entailed fine-tuning a semi-supervised concept extraction model by incorporating weight variables for hate target classification. Additionally, the author implemented a domain adaptation phase to detect targets and associated aspects in both the 'sexism' and 'racism' categories of the hate speech dataset.

2.3 Textual Stance Classification (TSC)

Various studies have delved into the classification of stance in text data across different domains, driven by the necessity to comprehend the dynamics within specific contexts, movements, and issues. The author (Upadhyaya et al., 2023) introduced MEMOCLiC, a multimodal multitasking framework for comprehensive stance detection in tweets. MEMOCLiC utilizes diverse embedding techniques and attention frameworks, incorporating learned emotional and offensive expressions. With a primary focus on stance detection, there were secondary tasks including emotion recognition and offensive language identification. The author's evaluation on climate change and benchmark

datasets highlights a notable $f1$ score of 93.76%.

In this TSC scheme, another study (Vaid et al., 2022) focused on addressing climate change concerns through the development of a stance detection and fine-grained classification system for related social media text. The study delved into linguistic features using part-of-speech tagging and named entity recognition. Two English datasets, ClimateStance and ClimateEng, each containing 3,777 annotated tweets, were introduced. State-of-the-art models like BERT, RoBERTa, and Distil-BERT are utilized for benchmarking.

3 Task and Dataset Description

The shared task encompasses three distinct sub-tasks: sub-task A, focusing on hate speech detection; sub-task B, centered around target detection; and sub-task C, concentrating on stance detection (Thapa et al., 2024). The organizers introduced a dataset called ClimaConvo (Shiwakoti et al., 2024), comprising 15,309 tweets related to various climate movements. Sub-tasks A, B, and C utilized subsets of this dataset.

3.1 Sub-Task A: Hate Speech Detection

This problem involves binary classification with two annotated labels: 'hate' and 'non-hate'. The dataset comprises a total of 7,284 training samples, 1,561 validation samples, and 1,562 test samples. The labels were encoded to 1 ('non-hate') and 2 ('hate').

3.2 Sub-Task B: Target Detection

Sub-task B is specifically focused on identifying targets in hate speech. The dataset dedicated to this sub-task consists of 699 training samples, along with 150 samples each for validation and testing. There are three classes in this dataset, these are 'individual', 'organization', and 'community'. The labels were encoded to 1 ('individual'), 2 ('organization'), and 3 ('community').

3.3 Sub-Task C: Stance Detection

The last sub-task revolves around identifying the stance in a given text, classifying it as 'support', 'oppose', and 'neutral'. This is particularly valuable for discerning whether activism is being supported or opposed by individuals. The dataset for sub-task C comprises of 7,284 training samples, 1,561 validation samples, and 1,562 test samples. The labels were encoded to 1 ('support'), 2 ('oppose'), and 3 ('neutral').

However, the dataset details are presented in Tables 1 and 2.

Tasks	Class	Initial	Duplicate Samples Removal	After sampling
Task A	1	6,385	5,899	5,899
	2	899	543	4,000
Task B	1	563	61	105
	2	105	105	105
	3	31	31	105
Task C	1	4,328	4,105	4,328
	2	700	190	2,000
	3	2,256	2,115	4,105

Table 1: Number of training samples per class after oversampling, considering the initial distribution and subsequent removal of duplicate entries.

4 Methodology

In this section, we delineate our methodology step by step. Figure 1 depicts a visual representation of the methodology.

4.1 Preprocessing of Data

Initially, we cleaned the provided dataset for all three sets—training, validation, and test. Employing a manually defined procedure using the Python regular expression library 're', we removed URLs, emojis, digits, and punctuation from the text. After that, we employed spaCy's [3] lemmatization by utilizing the English language model 'en_core_web_sm'. Considering that stopwords may not always be essential for classification and given the higher average length of the text, we removed stopwords using NLTK's [4] package 'stopwords' (Jefriyanto et al., 2023).

4.2 Duplicate Samples Removal from Dataset

To strengthen the instances of class 'hate' in sub-task A, samples from sub-task B were combined with sub-task A, labeling them as 'hate'. It was possible to do so because all the samples in sub-task B correspond to hate tweets targeting a specific audience. Samples of 'hate' class increased to 1,898, while 'non-hate' class samples remained at 6,385 after concatenation. However, the sub-tasks A, B, and C contain 1,008, 49, and 874 duplicate samples, which were removed eventually.

[3] https://spacy.io/

[4] https://www.nltk.org/

Task	Class	Train				Dev				Test			
		SC	TW	UW	AL	SC	TW	UW	AL	SC	TW	UW	AL
Task A	1	5,899	10,343	12,521	155	1,371	23,820	5,178	17	1,374	23,603	5,278	17
	2	543	10,211	3,157		190	2,962	970		188	3,171	1,078	
Task B	1	61	1,213	738	169	120	1,595	261	15	121	1,573	200	15
	2	105	2,166	1,142		23	472	330		23	500	367	
	3	31	588	407		7	181	154		6	130	112	
Task C	1	4,105	74,452	10,513	156	897	16,365	4,226	18	921	16,364	4,238	17
	2	190	3,530	1,657		153	2,177	587		141	2,005	538	
	3	2,115	37,360	7,463		511	88,857	3,048		500	8,405	2,772	

Table 2: Overall statistics of the dataset after the removal of duplicate entries. Here, SC, TW, UW, and AL denote sample count, total words, unique words, and average length, respectively.

4.3 Data Oversampling

This section is crucial as all tasks face a class imbalance issue, requiring an effective class distribution handling strategy for an improved $f1$ score. In all the sub-tasks, random oversampling (Gosain and Sardana, 2017) was employed to address the imbalance and enhance the model's ability to learn minority class patterns. While doing oversampling, careful consideration was given to the class distribution scenario after duplicate samples removal. It ensured a balanced approach by not oversampling a particular class too much, especially one with a very low distribution, and avoided the potential loss of focus on the majority class. The number of training samples after oversampling is provided in Table 1.

4.4 Extraction of Features

We employed various feature extraction methods, namely TF-IDF and Word2Vec for machine learning, fastText and GloVe for deep learning models.

TF-IDF is a numerical statistic indicating the importance of a term within a document relative to its occurrence across the entire dataset. For TF-IDF, we employed the default character n-gram as the analyzer.

Word2Vec embeddings (Mikolov et al., 2013) were generated using the 'en_core_web_sm' model in spaCy. Word2Vec is a popular technique for mapping words to dense vectors in a continuous vector space.

fastText embeddings with 300 dimensions were used for training DL models. fastText, an extension of Word2Vec, represents words as bags of character n-grams, enabling it to capture subword information, especially effective for morphologically rich languages and handling out-of-vocabulary words (Bojanowski et al., 2017).

GloVe constructs word vectors based on global statistical information of word co-occurrences across the entire corpus, capturing comprehensive semantic relationships for word meanings (Pennington et al., 2014). 'Glove.twitter.27B.100d' model was utilized as GloVe embedding, leveraging 100-dimensional word embeddings.

4.5 Machine Learning Models

Our exploration into ML model selection commenced with the consideration of four prominent models: RF, LR, SVM, and Multinomial Naive Bayes (MNB) (Sarker, 2021). These models have demonstrated superior performance in text classification tasks, motivating our choice. However, identifying optimal hyperparameters is critical, given their substantial impact on model performance. To address this challenge, we conducted a systematic search to determine the most suitable parameters for each model.

Model	Hyper-parameters
RF	n_estimators = 1000, min_samples_split = 2 min_samples_leaf = 1
MNB	alpha = 0.1, fit_prior = true, class_prior = false
SVM	C = 1, kernel = 'linear'
LR	solver = 'liblinear', penalty = 'l2'

Table 3: ML model's hyperparameter setting.

4.6 Deep Learning Models

In the development of text classification models, diverse deep learning architectures were investigated to tackle the intricacies of the task.

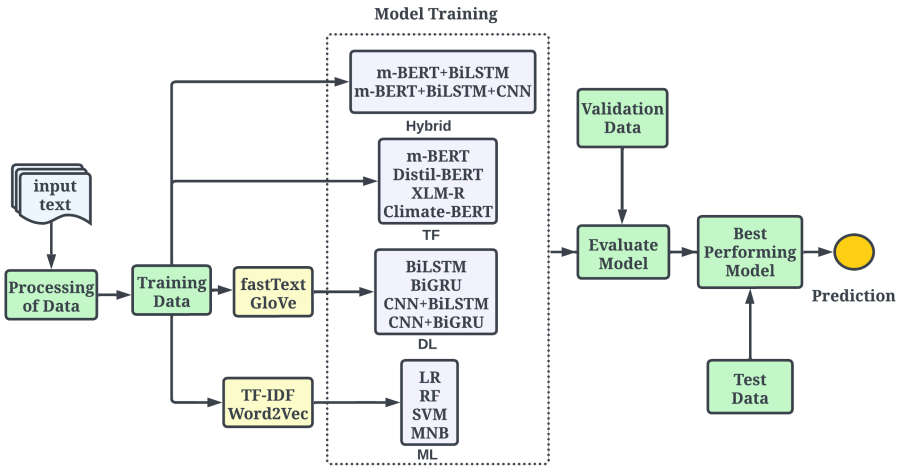

Figure 1: Visual representation of methodology.

BiLSTM: The initial model, employing a Bidirectional Long Short-Term Memory (**BiLSTM**) (Kalchbrenner et al., 2015) layer, served as the foundation. It featured a 100-dimensional embedding layer initialized with pre-trained word embeddings, a BiLSTM layer with 64 units for sequential data processing, followed by flattening and dense layers with dropout for regularization. This architecture laid the groundwork for subsequent models.

BiLSTM+CNN: The second model expanded on the BiLSTM design by integrating Convolutional Neural Network (CNN) components to make a hybrid BiLSTM+CNN model (Gehring et al., 2017). Additional Conv1D and MaxPooling1D layers were introduced to capture local features, enhancing the model's ability to discern patterns within the data.

CNN+GRU: The third model adopted another hybrid approach, combining CNN and Gated Recurrent Unit (GRU) layers to make CNN+GRU (Gehring et al., 2016). A Conv1D layer with 128 filters and a kernel size of 5 was followed by max-pooling, enhancing feature extraction. The bidirectional GRU (BiGRU) layer with 64 units provided a nuanced understanding of sequential dependencies. The model incorporated dense layers with dropout for regularization and concluded with an output layer.

BiGRU: The final model leveraged Bidirectional GRU (Cho et al., 2014) layers exclusively. It featured a 300-dimensional embedding layer, BiGRU with 256 units, and subsequent dense layers leading to an output layer. All the models underwent some common hyperparameters, which are shown in Table 4.

Parameters	Value
Learning Rate	$1e^{-3}$
Optimizer	Adam
Batch Size	32
AF(Hidden Layer)	Relu
AF(Output Layer)	Sigmoid (task A) Softmax (task B & C)
Dropout Rate	0.2

Table 4: DL model's hyperparameter setting, AF denotes the Activation Function.

4.7 Transformer-based Models

We conducted experiments using four pre-trained transformer-based models: m-BERT (Devlin et al., 2019), Distil-BERT (Sanh et al., 2019), XLM-R (Conneau et al., 2020), and Climate-BERT (Webersinke et al., 2021). To optimize training, we leveraged the 'fitonecycle' method from the ktrain library (Maiya, 2022). Prior to model training, we employed the 'find' method to visualize the learning rate curve, aiding in the identification of the optimal learning rate for each transformer-based model. Consequently, the learning rates and epochs varied among the models. Due to the substantial

Models	LR	Epochs	Batch Size	Max Length
m-BERT	$3e^{-5}$	10	16	256
Distil-BERT	$3e^{-5}$	12	16	
XLM-R	$2e^{-5}$	10	8	
Climate-BERT	$3e^{-5}$	10	16	

Table 5: Transformer-based model's hyperparameter setting. Here LR means Learning Rate.

volume of words and text size in tasks A and B, we adjusted the batch size accordingly, particularly for models such as XLM-R, ensuring efficient processing of the extensive textual data. We imported the transformer-based models from the 'Hugging Face' (Wolf et al., 2019). The detailed parameter settings are given in Table 5.

4.8 Hybrid Models

We experimented with BERT embedding by proposing two hybrid models. We considered m-BERT+BiLSTM (Jia, 2023) and m-BERT+BiLSTM+CNN (Mustavi Maheen et al., 2022) models. Figure 2 shows the overview of the hybrid models for both BiLSTM and BiLSTM+CNN utilizing BERT embedding.

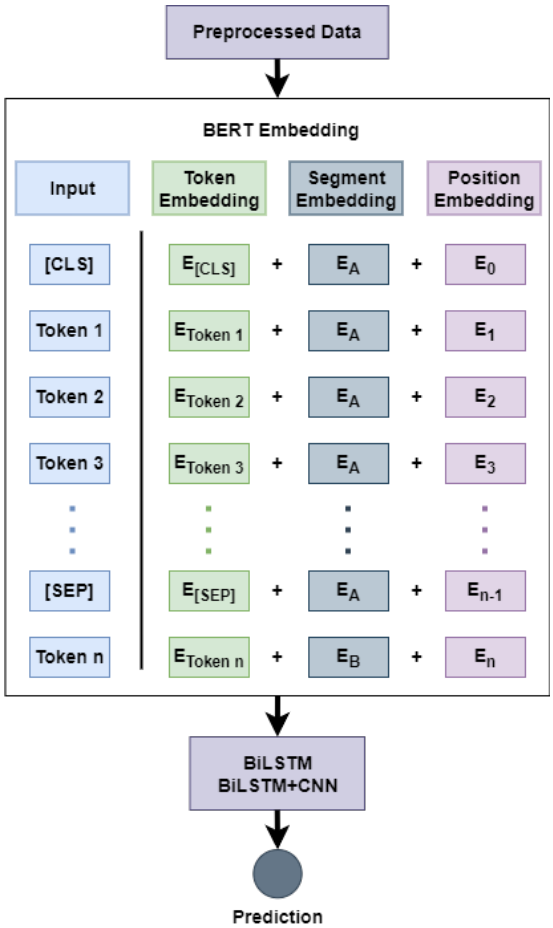

Figure 2: Overview of hybrid models.

m-BERT+BiLSTM: The first model integrates BERT embeddings, Bidirectional LSTM, and pooling layers for text classification. BERT embeddings are subject to dropout regularization and reshaped into a 3D tensor. A Bidirectional LSTM layer captures sequential context, while global pooling extracts key features. These pooled outputs are concatenated and processed through dense layers with ReLU activation and dropout. The final layer utilizes an activation function for ultimate prediction. This architecture leverages BERT's contextual embeddings and Bidirectional LSTM's sequential learning for enhanced text classification.

m-BERT+BiLSTM+CNN: The second model, combines BERT embeddings, Bidirectional LSTM, and a Convolutional Neural Network (CNN) to capture diverse contextual and sequential patterns in the input text. BERT embeddings undergo dropout regularization, followed by reshaping and processing through a bidirectional LSTM and a 1D CNN layer. Global average pooling, global max pooling, and flattened CNN outputs are concatenated. Two dense layers with dropout provide additional abstraction, leading to an output layer. This architecture aims to leverage the strengths of BERT embeddings, LSTM, and CNN to enhance the model's ability to discern patterns in sequential data for accurate classification. The parameter setting remains the same as the parameter settings for DL models (see Table 4).

5 Results and Analysis

In this section, we delve into a comprehensive comparative analysis of our proposed models across all three sub-tasks. Table 6 presents such a comprehensive evaluation.

5.1 Sub-Task A

In sub-task A, RF with Word2Vec demonstrated superior efficiency in achieving a higher $f1$ score compared to the TF-IDF counterpart. It outperformed all other ML models with a notable $f1$ score of 0.89. Even though several ML models performed almost nearly well, the MNB appeared to perform poorly on non-oversampled data. MNB struggled to handle class imbalance and due to the lack of minority class instances ('hate'), it is classifying all the samples into 'non-hate'. Among DL models, the hybrid CNN+BiGRU with GloVe embedding attained an impressive $f1$ score of 0.91 even before oversampling. As GloVe utilized global statistical information by offering improved representation of word meanings, the CNN+BiGRU model took benefit of this. It also performed better with fastText embedding as well. For transformer-based models, m-BERT excelled with a $f1$ score of 0.91, which was similar to CNN+BiGRU (GloVe). Its performance before

FET	Models	Without Oversampling						With Oversampling					
		Task A		Task B		Task C		Task A		Task B		Task C	
		$f1$	Acc	$f1$	Acc	$f1$	Acc	$f1$	Acc	$f1$	Acc	$f1$	Acc
TF-IDF	RF	0.81	0.91	0.56	0.88	**0.67**	0.69	0.80	0.92	**0.69**	0.89	0.28	0.68
	LR	0.83	0.91	**0.65**	0.91	0.65	0.91	0.85	0.93	0.59	0.87	**0.39**	0.64
	SVM	**0.86**	0.95	0.63	0.89	0.63	0.89	**0.88**	0.95	0.63	0.89	**0.39**	0.63
	MNB	0.47	0.88	0.56	0.88	0.56	0.88	0.83	0.91	0.66	0.89	0.34	0.62
Word2Vec	RF	**0.89**	0.95	0.54	0.87	0.54	0.87	**0.88**	0.94	0.67	0.86	0.53	0.64
	LR	0.73	0.83	0.70	0.89	**0.67**	0.88	0.72	0.83	0.70	0.89	0.55	0.57
	SVM	0.86	0.95	**0.71**	0.89	**0.67**	0.89	0.72	0.83	**0.71**	0.89	**0.56**	0.59
	MNB	0.47	0.87	0.54	0.87	0.55	0.87	0.71	0.84	**0.71**	0.89	0.27	0.60
GloVe	BiLSTM	0.87	0.95	**0.61**	0.87	0.65	0.65	0.47	0.88	**0.63**	0.86	0.66	0.67
	BiGRU	0.90	0.96	0.53	0.88	**0.66**	0.69	**0.80**	0.89	**0.63**	0.89	**0.67**	0.68
	BiLSTM+CNN	0.87	0.95	0.58	0.88	0.64	0.65	0.47	0.88	0.57	0.85	0.63	0.63
	CNN+BiGRU	**0.91**	0.96	**0.61**	0.87	0.59	0.67	0.47	0.88	0.51	0.82	0.66	0.68
fastText	BiLSTM	0.56	0.86	0.54	0.83	0.56	0.61	0.56	0.86	0.56	0.87	0.64	0.65
	BiGRU	0.70	0.81	0.57	0.85	0.60	0.61	0.70	0.81	0.59	0.87	0.64	0.64
	BiLSTM+CNN	0.85	0.92	**0.62**	0.88	0.63	0.65	0.84	0.92	**0.68**	0.87	**0.66**	0.66
	CNN+BiGRU	**0.90**	0.95	0.59	0.83	**0.64**	0.67	**0.90**	0.95	0.65	0.88	**0.66**	0.66
	m-BERT	**0.91**	0.96	0.64	0.86	0.63	0.62	**0.91**	0.96	**0.74**	0.89	0.66	0.65
	Distil-BERT	0.88	0.95	**0.65**	0.85	0.62	0.64	0.86	0.94	**0.74**	0.89	**0.67**	0.65
	XLM-R	0.82	0.93	0.63	0.85	0.60	0.62	0.88	0.88	0.70	0.88	0.65	0.69
	Climate-BERT	0.90	0.96	0.63	0.88	**0.67**	0.71	**0.91**	0.96	0.71	0.89	**0.67**	0.68
	m-BERT+BiLSTM	**0.83**	0.94	**0.54**	0.87	0.25	0.59	**0.73**	0.85	0.53	0.86	0.25	0.59
	m-BERT+BiLSTM+CNN	0.66	0.77	0.48	0.82	**0.31**	0.61	0.31	0.32	0.50	0.85	**0.62**	0.62

Table 6: Result comparison over test data. Here FET means feature extraction technique, $f1$ denotes macro-averaged $f1$ score and Acc means Accuracy.

and after oversampling remains the same. Finally, m-BERT and CNN+BiGRU (GloVe) embedding were identified as the best-performing models for this sub-task.

5.2 Sub-Task B

Turning to the sub-task B, m-BERT and Distil-BERT exhibited identical $f1$ scores of 0.74 in the oversampled dataset. Which suggests a very crucial improvement after increasing minority classes. Due to the increased number of samples, the BERT models were able to effectively identify the semantic and contextual meaning of the tweets rigorously. But interestingly the hybrid model with BERT embedding underperformed, even trailing behind some ML and DL models. The BERT's complex pre-trained architecture didn't provide substantial benefits compared to other embeddings like GloVe and fastText. ML models showed improved performance after oversampling. SVM and MNB achieved a $f1$ score of 0.71 in the oversampled dataset with Word2Vec embedding. DL models like BiLSTM and BiGRU with GloVe embedding performed better on oversampled data compared to non-oversampled counterparts. However, BiLSTM+CNN with fastText embedding appeared to be the best-performing DL model with a $f1$ score of 0.68. Consequently, m-BERT and Distil-BERT were identified as the best models for this sub-task. We submitted all the models for the shared task and finalized m-BERT for the final leaderboard standings.

5.3 Sub-Task C

In the case of ML models, it is seen that the performance of ML models on oversampled data degraded significantly. The reason is that the heavily imbalanced dataset along with the two most challenging and confusing classes 'support' and 'neutral' made classification difficult. The confusion of the classification was further fueled by oversampled data, resulting in poor performance with TF-IDF and Word2Vec. Nevertheless, transformer-based models surpassed the baseline score, indicating promise. Climate-BERT consistently performed best with a $f1$ score of 0.67, on both oversampled and non-oversampled data. As it is heavily trained on climate-related texts, therefore oversampling didn't affect its performance in this case. On the other hand, hybrid models that utilized BERT embedding performed better in oversampled data. Because of the capability to handle larger datasets, the BERT embedding appeared to perform better when dataset size increased by oversampling.

5.4 Performance Comparison

Table 7 shows that the performance of our team was promising as compared to other participating teams. In all of the sub-tasks, we were able to beat the baseline scores provided by the organizer on the ClimaConvo dataset.

Team Name	Sub-Task A				
	R	P	$f1$	Acc	Rank
CUET_Binary_Hackers	**0.9173**	**0.9116**	**0.9144**	**0.9635**	**1st**
AAST-NLP	0.8654	0.9231	0.8914	0.9571	2nd
MasonPerplexity	0.8689	0.9112	0.8885	0.9552	5th
Baseline Score	-	-	**0.708**	**0.901**	-
	Sub-task B				
MasonPerplexity	0.7823	0.8133	0.7858	0.9133	1st
AAST-NLP	0.7706	0.7689	0.7665	0.9133	3rd
CUET_Binary_Hackers	**0.7533**	**0.7431**	**0.7433**	**0.9000**	**6th**
Baseline Score	**-**	**-**	**0.716**	**0.901**	**-**
	Sub-Task C				
Hamison-Generative	0.7223	0.7827	0.7479	0.7478	1st
CUET_Binary_Hackers	**0.6691**	**0.6908**	**0.6794**	**0.6613**	**15th**
Z-AGI Labs	0.6294	0.7926	0.6372	0.6908	16th
Baseline Score	**-**	**-**	**0.651**	**0.545**	**-**

Table 7: Short rank list for all sub-tasks. P, R, $f1$, Acc denote precision, recall, macro $f1$ score, and accuracy respectively.

6 Error Analysis

The study investigated the performance of m-BERT (sub-task A and B) and Climate-BERT (sub-task C) models using quantitative and qualitative methods. Text samples were randomly chosen for all sub-tasks to facilitate quantitative analysis.

6.1 Sub-Task A

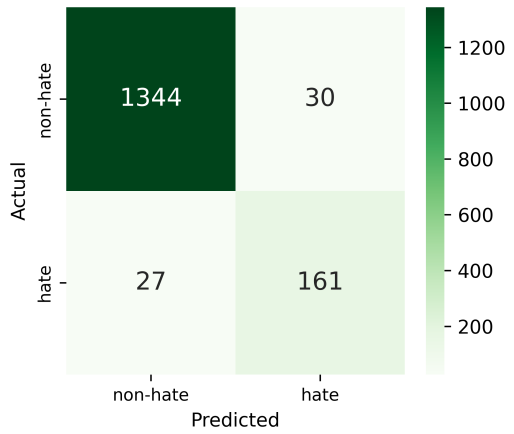

Figure 3: Confusion matrix for sub-task A by the m-BERT model.

Figure 3 indicates that out of 1,370 'non-hate' samples, 30 were misclassified, while 27 'hate' samples were misclassified as 'non-hate', despite oversampling achieving nearly 85% accuracy in 'hate' samples. The presence of common hashtags in most of the samples led to the misclassification of samples.

Table 8 describes the qualitative analysis of sub-task A, where samples 1, 2, and 3 were predicted the same as their actual label. However, samples 4, 5, and 6 resulted in misclassification by the m-BERT model.

Test Sample	Actual	Predicted
Sample 1: Love the artwork despite doubting its factual accuracy	non-hate	non-hate
Sample 2: Vladimir Putin is a global warming accelerationist. CdnNatSec FridaysForFuture	hate	hate
Sample 3: Happy EarthDay!	non-hate	non-hte
Sample 4: apparently now we have a "Planet Farm" nearby, guys!!climatechange ConsciousPlanet FridaysForFuture	non-hate	hate
Sample 5: Germany goes nuclear! Atomkraft NuclearPower FridaysForFuture Gruenen GruenerMist	non-hate	hate
Sample 6: Stop with the bullshit forecasts. @ExtinctionR ClimateStrike PeopleNotProfit FridaysForFuture 1BillionClimateVoices	hate	non-hate

Table 8: Some test samples for sub-task A, predicted by the m-BERT.

6.2 Sub-Task B

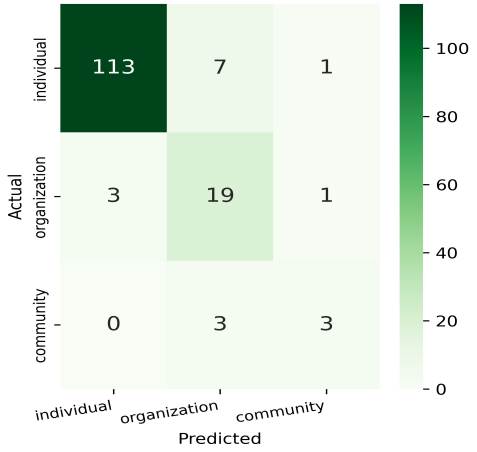

Figure 4: Confusion matrix for sub-task B by the m-BERT model.

Figure 4 reveals a higher misclassification rate in class 3 ('community') due to the lower number of training samples, resulting in a 50% misclassification rate. Classes 1 ('individual') and 2 ('organization') exhibited lower misclassification rates, with class 2 slightly higher due to class imbalance issues.

Qualitative analysis of sub-task B was presented in Table 9, where samples 1, 2, and 3 were misclas-

sified by the m-BERT model. However, samples 4, 5, and 6 were predicted correctly, matching the actual labels of the samples.

Test Sample	Actual	Predicted
Sample 1: @Citi spent the last 5 years investing $285 billion into destroying our futures. FridaysForFuture Divest	individual	organization
Sample 2: Vladimir Putin is a global warming accelerationist. CdnNatSec FridaysForFuture	individual	organization
Sample 3: If any politicians you encounter tomorrow have been reluctant about ClimateActionNow and/or providing Reparations for LossAndDamage, PLEASE trap them in a WallPinOfLove (or, in this case, confrontation)!!! GlobalClimateStrike FridaysForFuture PeopleNotProfit @GretaThunberg	community	organization
Sample 4: Fuck Greta not the planet savetheplanet FridaysForFuture	individual	individual
Sample 5: Elections matter. Stop electing climate deniers and fossil fuels industry puppets.PeopleNotProfit ActOnClimate Australia auspol ClimateCrisis ExtinctionRebellion environment FFF FridaysForFuture	organization	organization
Sample 6: @dw_environment @Luisamneubauer @Fridays4future has remained influenced by strong left ideology/persons and denies the science using (existing) nuclear in climate/independence policies.	community	community

Table 9: Some test samples for sub-task B, predicted by the m-BERT.

6.3 Sub-Task C

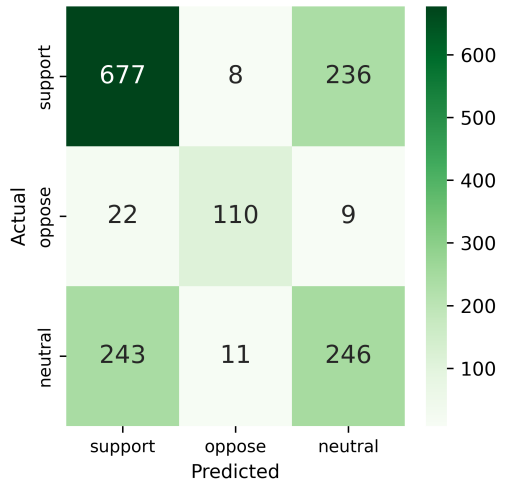

Figure 5: Confusion matrix for sub-task C by the Climate-BERT model.

Figure 5 illustrates misclassifications, particularly prominent between class 1 ('support') and class 3 ('neutral') in sub-task C. Among the predictions, 236 samples were classified as class 2 ('oppose'), while 243 were classified as class 3. The issue was exacerbated by class imbalance, resulting in 31 misclassified samples out of 141. The model struggled to differentiate between classes 1 and 3 due to their proximity.

Table 10 presents the output of several sample texts analyzed by the Climate-BERT model. Samples 1, 2, 3, and 4 were predicted dissimilar to their actual labels, whereas samples 5, 6, and 7 were predicted correctly, aligning with the actual labels.

Test Sample	Actual	Predicted
Sample 1: 4 year of FridaysForFuture	neutral	support
Sample 2: Gretas Gamlingar stockholm FridaysForFuture	neutral	oppose
Sample 3: Fuck Greta not the planet savetheplanet FridaysForFuture	oppose	support
Sample 4: Education is a human right! FridaysForFuture EducateGirlsForClimateJustice	support	neutral
Sample 5: Love and kindness are never wasted. KindnessMatters FridaysForFuture GlobalGoals	support	support
Sample 6: Germany goes nuclear! Atomkraft NuclearPower FridaysForFuture Gruenen GruenerMist	oppose	oppose
Sample 7: Is anything more dangerous than ClimateCrisis? FridaysForFuture	neutral	neutral

Table 10: Some test samples for sub-task C, predicted by the Climate-BERT model.

7 Conclusion

In this paper, we present a fine-tuned approach utilizing various models, specifically proposing fine-tuned m-BERT, Distil-BERT, Climate-BERT, and CNN+BiGRU. The results indicate that m-BERT achieved a higher $f1$ score for both sub-tasks A and B. The highest $f1$ score that we achieved for sub-task A is 0.91, for sub-task B it is 0.74, and for sub-task C it is 0.67. Several models like Climate-BERT, BiGRU, LR, and SVM performed equally well with the same $f1$ score for sub-task C. Our paper includes a detailed comparison among several models, both before and after addressing the class imbalance in the datasets. Notably, in most cases, the performance showed significant improvement. This paper also delved into effective preprocessing of data and data oversampling. These findings will create new opportunities for upcoming research work, drawing inspiration from this paper.

Limitations

Our system exhibits some key limitations:

- The significance and novelty of the research findings could be increased by introducing novel models or approaches.
- The efficiency of imbalance handling in detection models can be increased by including a wider range of data augmentation approaches.

References

Sergio Arce-García, Jesús Díaz-Campo, and Belén Cambronero-Saiz. 2023. Online hate speech and emotions on Twitter: a case study of Greta Thunberg at the UN Climate Change Conference COP25 in 2019. *Social Network Analysis and Mining*, 13(1):48.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Alexis Conneau, Kartik 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Benjamin Damoah, Sagini Keengwe, Samuel Owusu, Clement Yeboah, and Francis Kekessie. 2023. The Global Climate and Environmental Protest: Student Environmental Activism a Transformative Defiance. *International Journal of Environmental, Sustainability, and Social Science*, 4(4):1180–1198.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *International Conference on Web and Social Media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Anjana Gosain and Saanchi Sardana. 2017. Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)*, pages 79–85. IEEE.

Neil Gunningham. 2019. Averting climate catastrophe: environmental activism, extinction rebellion and coalitions of influence. *King's Law Journal*, 30(2):194–202.

Jefriyanto Jefriyanto, Nur Ainun, and Muchamad Arif Al Ardha. 2023. Application of Naïve Bayes Classification to Analyze Performance Using Stopwords. *Journal of Information System, Technology and Engineering*, 1(2):49–53.

Tao Jia. 2023. A Named Entity Recognition Method Based on Pre trained Models MBERT and BiLSTM. In *Proceedings of the 2023 6th International Conference on Information Science and Systems*, pages 30–35.

Nal Kalchbrenner, Ivo Danihelka, and Alex Graves. 2015. Grid long short-term memory. *arXiv preprint arXiv:1507.01526*.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15.

Sandy Kurniawan and Indra Budi. 2020. Indonesian tweets hate speech target classification using machine learning. In *2020 Fifth International Conference on Informatics and Computing (ICIC)*, pages 1–5. IEEE.

Thomas Laux. 2021. What makes a global movement? Analyzing the conditions for strong participation in the climate strike. *Social Science Information*, 60(3):413–435.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Arun S Maiya. 2022. ktrain: A low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1):7070–7075.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Alexey Mikhaylov, Nikita Moiseev, Kirill Aleshin, and Thomas Burkhardt. 2020. Global climate change and greenhouse effect. *Entrepreneurship and Sustainability Issues*, 7(4):2897.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

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.

Sally Neas, Ann Ward, and Benjamin Bowman. 2022. Young people's climate activism: A review of the literature. *Frontiers in Political Science*, 4:940876.

Parihar, Anil Singh and Thapa, Surendrabikram and Mishra, Sushruti. 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.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics.

Iqbal H Sarker. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

David Schlosberg and Lisette B Collins. 2014. From environmental to climate justice: climate change and the discourse of environmental justice. *Wiley Interdisciplinary Reviews: Climate Change*, 5(3):359–374.

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*.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

Viktoria Spaiser, Nicole Nisbett, and Cristina G Stefan. 2022. "How dare you?"—The normative challenge posed by Fridays for Future. *PLOS Climate*, 1(10):e0000053.

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)*.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A Multi-task Model for Emotion and Offensive Aided Stance Detection of Climate Change Tweets. In *Proceedings of the ACM Web Conference 2023*, pages 3948–3958.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

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*.

Sarah R Weiskopf, Madeleine A Rubenstein, Lisa G Crozier, Sarah Gaichas, Roger Griffis, Jessica E Halofsky, Kimberly JW Hyde, Toni Lyn Morelli, Jeffrey T Morisette, Roldan C Muñoz, et al. 2020. Climate change effects on biodiversity, ecosystems, ecosystem services, and natural resource management in the United States. *Science of the Total Environment*, 733:137782.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.