IUST at ClimateActivism 2024: Towards Optimal Stance Detection: A Systematic Study of Architectural Choices and Data Cleaning Techniques

Ghazaleh Mahmoudi and Sauleh Eetemadi

School of Computer Engineering, Iran University of Science and Technology, Iran

Introduction

Climate change is a pressing issue that affects ecosystems, economies, and communities globally. Understanding public perspectives on this crucial topic is more important than ever. Leveraging Natural Language Processing (NLP) techniques to analyze public stance toward climate change using Twitter data offers a novel approach to capturing diverse viewpoints in real-time.

In this context, the ClimateActivism 2024 Shared Task introduces three sub-tasks focused on Stance and Hate Event Detection. Our research delves into improving stance detection models on climate change-related tweets by exploring various model architectures, data cleaning methods, and addressing data imbalance through augmentation techniques. By combining Convolutional Neural Networks (CNN) and BERTweet with Weighted Cross Entropy as the loss function, we demonstrate superior performance compared to Feedforward Neural Networks (FNNs). Our findings underscore the importance of data augmentation and effective data cleaning methods, such as removing URLs and usernames, in enhancing model accuracy for stance detection in climate change discourse on Twitter.

Data

Data Preprocessing

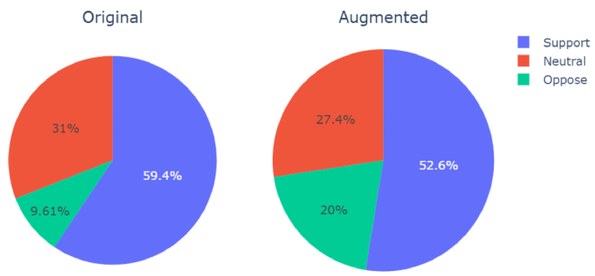
We define 7 level of data preprocessing. The defined methods will involve increasing levels of text input cleaning, from the least to the most aggressive.

- C1: **Original Tweet**
- C2: **Removing URL**
- C3: **Removing Username**
- C4: **Removing URL and Username**
- C5: **Removing URL and Username and split hashtag.** For example #FridaysForFuture becomes Fridays For Future.
- C6: **Removing URL and Username and split hashtag and lower case**
- C7: **Complete Cleaning.** Contains removing URL, username, stop words, punctuation, converting all letters to lowercase, and split hashtag.

Data Augmentation

To address data imbalance challenge we two different methods to generate additional data.

- **Substitution**: use synonym substitution as an augmentation method.
- **Round-trip translation**: translate the English texts to German and then back, to generate extra data.



Methodology

The study proposed a model with four modules and conducted multiple experiments to find the best parameters for each module.

The Optuna library was used to select the optimal model configuration based on the Macro F1-score. The search space for each module was defined to find the most suitable values.

- **Embedding**
- **Classifier**
- **Optimizer**
- **Loss Function**

Parameter	Search Space
Embedding	BERT, RoBERTa, BERTweet, XLM-RoBERTa
Classifier	CNN, FNN
Optimizer	Adam, AdamW, RMSPropb, SGD
Loss Function	Weighted Cross Entropy, Focal

Architecture search space

Hyper-parameters used in training stages are selected via tuning using the Optuna library. We choose the optimal hyperparameters by the Macro F1-score on the development set.

Hyperparameter	Search Space
Dropout	[0.1: 0.5]
Learning Rate	[0.00001: 0.01]
Batch Size	[4, 8]
Focal Gamma	[1, 2, 3, 4, 5]

Hyperparameters search space

Results

We tested 14 configurations in total, including 7 modes for data cleaning and 2 modes for input data. For each configuration, we selected model. parameters and hyperparameters using Optuna and performed fine-tuning for 20 trials. In each trial, the parameters are selected using the sampling method TPE, based on the defined search space

	Parameter	Value
Model Configuration	Epoch	8
	Batch Size	4
	Dropout	0.5
	Learning Rate	0.007903
	Learning Scheduler	Linear Scheduler with Warmup
Hyperparameters	Embedding	BERTweet
	Classifier	CNN
	Optimizer	SGD
	Loss Function	Weighted Cross Entropy

Best model configuration and hyperparameters.

The results indicate that C3(Removing username) and C4 (Removing URL and username) are significantly better than C1(Original Tweet Text) and C7(Complete cleaning). Thus, the influence of data cleaning methods on the final results is clearly evident.

Cleaning	F1-Score
C1	73.98±0.0012
C2	73.92±0.0017
C3	74.35±0.0015†
C4	74.11±0.0029†
C5	73.76±0.0014
C6	73.72±0.0009
C7	72.42±0.0020

Data Cleaning impact

By repeating the experiment with the best configuration and only changed the classifier, it demonstrated the superiority of CNN over FNN.

Classifier	Cleaning	F1-Score
CNN	C3	74.35±0.0015†
	C4	74.11±0.0029†
FNN	C3	73.73±0.0058
	C4	73.91±0.0045

Classifier impact

Conclusion

This work involved a systematic exploration of model architecture and data cleaning methods. We find that the optimal configuration combining BERTweet and CNN with Weighted Cross Entropy and SGD, along with data augmentation. We demonstrate that a combination of CNN and Encoder-only models such as BERTweet outperforms FNNs. Moreover, by utilizing data augmentation, we are able to overcome the challenge of data imbalance. Our best system achieves **74.47%** F1- Score on the unseen test set, outperforming the **baseline by 19.97% and ranked 3th among 19 participants**.