HAMiSoN-Ensemble at ClimateActivism 2024: Ensemble of RoBERTa, Llama 2, and Multi-task for Stance Detection

Raquel Rodriguez-Garcia and **Julio Reyes-Montesinos** and
Jesús M. Fraile-Hernandez and **Anselmo Peñas**
NLP & IR Group
UNED, Spain
{rrodriguez, jreyes, jfraile, anselmo}@lsi.uned.es

Abstract

CASE @ EACL 2024 proposes a shared task on Stance and Hate Event Detection for Climate Activism discourse. For our participation in the stance detection task, we propose an ensemble of different approaches: a transformer-based model (RoBERTa), a generative Large Language Model (Llama 2), and a Multi-Task Learning model. Our main goal is twofold: to study the effect of augmenting the training data with external datasets, and to examine the contribution of several, diverse models through a voting ensemble. The results show that if we take the best configuration during training for each of the three models (RoBERTa, Llama 2 and MTL), the ensemble would have ranked first with the highest F_1 on the leaderboard for the stance detection subtask.

1 Introduction

Social media is a popular tool and has adopted an essential role in this day and age. With its massive spread and usage, a global discourse arises regarding numerous topics. Climate change has become a most prominent topic, as well as a very polarized one (Tyagi et al., 2020; Chen et al., 2023). As debates develop, the emergence of hate speech becomes a concern that must not be left unattended.

Although the case for freedom of speech has been voiced, it cannot be confused with a complete lack of regulations. Touting that rhetoric has brought harmful effects (Hickey et al., 2023). It delves into the paradox of tolerance, where unlimited freedom of speech can cause the corrosion of our society. Rampant hate speech can create a breeding ground for intolerance and discrimination. All of these circumstances justify the necessity to study controversial public discourse, to promote online safety and inclusion.

For the reasons argued above, the Climate Activism Stance and Hate Event Detection task at CASE 2024 (Thapa et al., 2024) has significant relevance. This task focuses on climate activism discourse, and it consists of three distinct subtasks: hate speech detection, target identification and stance detection. It can provide valuable knowledge on the diffusion of hate speech and the polarization of users' stances, addressing some current open challenges (Parihar et al., 2021).

As described in this paper, our proposal leverages an ensemble voting system with two different voting strategies for the stance detection subtask. Ensemble voting has been used in other stance detection shared tasks (Cignarella et al., 2020), achieving the best results. These systems provide some additional advantages, such as model regularization and an increase of diversity (Polikar, 2006), as they consider different approaches simultaneously. For our ensemble, we exploit three different systems: a transformer-based baseline model, a Large Language Model and Multi-Task Learning.

Beyond exploring a set of diverse systems for the proposed task, our approach has the goal of studying the effect of external data on the stance detection subtask. We aim to determine the effect that external training data has on our proposed models, and to evaluate the suitability of these external datasets towards improving a model's performance in the context of climate activism. To this end, we propose two datasets related to hate speech and stance detection that we detail below.

This paper is organized as follows. In section 2 we introduce the dataset for the task. In section 3 we present the strategy for the ensemble models, as well as the additional data that were used. In section 4 we introduce our results, we discuss them in section 5, and we perform a post-competition analysis in section 6. Finally, in section 7, we exhibit our conclusions and future work.

2 Dataset and Task

This shared task, Climate Activism Stance and Hate Event Detection, uses the ClimaConvo dataset in-

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 118–124
March 22, 2024 ©2024 Association for Computational Linguistics

troduced in Shiwakoti et al. (2024). It consists of tweets containing hashtags from a curated list linked to climate change and climate activism, collected over a one-year period. Non-English tweets were filtered out. The final dataset only reflects the textual content of the tweets and was manually annotated in six dimensions. The shared task at hand is based on a subset of ClimaConvo and contains 10,407 instances. Below, we describe the three subtasks proposed over this dataset.

2.1 Subtask A

The goal of subtask A is to establish whether a tweet contains hate speech or not. This is a binary classification task with HATE SPEECH and NO HATE SPEECH as the annotated labels.

2.2 Subtask B

Subtask B aims to discover the target of the hate speech, with a multiclass classification task with the INDIVIDUAL, ORGANIZATION, or COMMUNITY labels. Subtask B is based on a smaller subset of 999 instances, corresponding to tweets where hate speech is present and labeled as DIRECTED in ClimaConvo (Shiwakoti et al., 2024).

2.3 Subtask C

Finally, the objective of subtask C is to determine the stance of the tweets. The data used for this task is the same as subtask A. Similarly to subtask B, this is a multi-class classification task with three labels: SUPPORT, OPPOSE and NEUTRAL.

3 Methodology

Different models have been employed for the ensemble described in this paper. In this section we review the external datasets used by the models, the pre-processing step applied to all the data sources, the descriptions of each model and the characteristics of the ensemble classifier. We aim at determining whether an ensemble makes a robust model, and whether the additional context of other datasets provides an advantage to this task.

3.1 External Data

We experiment with two main data sources: an offensive language and target dataset, and a stance dataset. Although we have only participated in subtask C with this ensemble, additional related data, as well as the hate speech and target subtasks, have been included in some of these models.

One of the considered data sources has been the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a), which was used in the SemEval 2019 Task 6 (Zampieri et al., 2019b). It is composed of Twitter data with each tweet being annotated for three subtasks: offensive language identification (whether a tweet is offensive or not), characterization of offense types (whether it is targeted or not) and offense target identification (the target of the offense: INDIVIDUAL, GROUP or OTHERS). The train and test sets have been combined for training, generating a total of 14,100 annotated samples for the offensive language identification and 4,089 for the target task.

In addition to the OLID dataset, the stance dataset by Mohammad et al. (2016a), used in the SemEval 2016 Task 6 (Mohammad et al., 2016b), has been included. This Twitter dataset is comprised of different sections, determined by the topics of the tweets. There is a total of 4,163 tweets organized by the topics of abortion, climate, Hillary Clinton, feminism and atheism. This is a multi-class classification task, which considers three classes: AGAINST, FAVOR or NONE. Using the same approach as with the previous dataset, the train and test sets have been combined.

3.2 Dataset Preparation

All our models use the text of the tweet as input. We pre-processed this text with a pipeline consisting of the following steps:

- Removal of URLs from tweets.
- Replacement of username mentions by the generic token `@USER`.
- Splitting of hashtags into individual words. To accomplish this endeavor we have utilized the Word Ninja[1] library, which uses a probabilistic division of concatenated words, based on the frequencies of unigrams from the English Wikipedia.

3.3 Model Description

For this ensemble, we leveraged three different approaches that participated individually in the shared task. Below, we discuss the characteristics of the models, as well as a description of each of the runs:

3.3.1 RoBERTa

We established baseline systems based on RoBERTa-base (Liu et al., 2019) transformers with

[1] https://github.com/keredson/wordninja

a classification head. We fine-tuned a RoBERTa-base model for each of the subtasks using only the data proposed in the shared task, and a second set of RoBERTa-base models on both the data proposed in the shared task and the additional data proposed for each subtask. With this, we aim at providing a baseline comparison of the impact of using additional training data in each subtask that subsequent models can elaborate on. A more in-detail description of our fine-tuning methodology for RoBERTa can be found in Reyes-Montesinos and Rodrigo (2024).

3.3.2 Llama 2

Next, we fine-tuned a Llama 2 7B Chat model with a final classification layer, using raw prompts. In this model, we start from the Llama 2 7B Chat model proposed by Meta in Touvron et al. (2023). Then, we removed the last linear layer to add another linear layer that has as input the last hidden state of the model and as output 3 neurons, one for each stance label. As model input, we use the tweets pre-processed as explained in 3.2, therefore not following the officially suggested tag format. Moreover, as it is a generative model, we have tested the zero-shot approach, but our low initial results led us to use the classification layer. The full description of the model and the zero-shot approach can be found in Fraile-Hernandez and Peñas (2024).

3.3.3 Multi-Task Learning

This approach leverages the potential of transformer-based Multi-Task Learning (MTL) for this subtask, and it is detailed at length in Rodriguez-Garcia and Centeno (2024). In our system, we implement a hard parameter sharing Multi-Task model, as was originally described by Caruana (1993). The model is composed of a shared RoBERTa encoder and one classification head for each different task the model is training for. Considering the capabilities of this approach to extract context from related information, some of our MTL models have been trained with the three subtasks: hate speech, target, and stance.

3.4 Ensemble Description

Two different approaches have been explored for the ensemble process, a majority voting strategy and a conservative strategy. In the majority voting, the predicted stance will be the majority of the votes of the three base models, and a tie is resolved by returning the NEUTRAL label, given that no consensus was reached between SUPPORT and OPPOSE.

In the conservative strategy, the predicted stance will be the label that is obtained by unanimity of the votes of the three base models. In the case of no unanimity for a label, this strategy would return the value NEUTRAL. This strategy was motivated due to the error analysis during validation. We observed that the models had problems correctly classifying the NEUTRAL label, and they tended to classify these instances as SUPPORT.

4 Results

In total, we performed 10 experiments: 4 ensembles and 6 individual component systems. Half of the runs were performed using only the CASE dataset and the other half using data additional to the CASE stance dataset. Specific hyperparameters and training details are reflected in the individual papers for each system.

For the CASE only runs, we fine-tune RoBERTa and Llama 2 models on only Subtask C data. The Multi-task Learning (MTL) system was fined tuned on subtask A, B and C data to fully extract the knowledge from the task.

Regarding the runs with additional external data, the RoBERTa systems use the climate only topic from the SemEval stance dataset, while the Llama 2 models make use of all the topics from that dataset. Finally, the Multitask Learning model adds only the offensive language identification and the target tasks from the OLID dataset.

Table 1 shows the F_1 macro value of the 4 different runs. The results are divided into **CASE** if only the CASE dataset has been used in the training, or **Added** if the models have been trained with the CASE dataset and the additional data. The results of the individual models used for the ensemble are also included, in addition the results of the Baseline model used in Shiwakoti et al. (2024). This model, named ClimateBERT (Webersinke et al., 2022), is an adaptation of a BERT model, a language model trained on a corpus sourced from climate-related news, abstracts, and reports. Furthermore, we compute an oracle to establish the upper limit of the ensemble. The ideal version of our systems predicts the correct class if any of the three components managed to predict it on its own.

	Approach	CASE	CASE + external
	Baseline	0.5450	
	Best model leaderboard	0.7483	
Indiv.	RoBERTa	**0.7495**	**0.7406**
	Llama 2	0.7366	0.7300
	MTL (submitted)	0.7295	0.7320
Ensemble	Conservative	0.7265	0.7287
	Majority vote	0.7479	0.7397
	Oracle (upper bound)	0.8332	0.8259

Table 1: Comparison of F_1-scores for the best submitted individual models and the ensembles constructed from them, both trained on only task data and task and external data.

5 Discussion

As noted in Table 1, the performance of our proposed models greatly surpasses the baseline proposed by the organizers. Our best ensemble model – using the majority voting strategy – comes up second on the leaderboard by F_1 score for Subtask C. Regarding the use of additional data, we see that performance only improves in MTL and worsens for both RoBERTa and LLama 2. In the case of Llama 2, it could be due to the fact that the external dataset we used covers several topics – only 13.5% of the instances were related to climate activism. This distribution of data can add noise to training. As for RoBERTa, we only used the additional data of the same topic. We conclude that the strategy of augmenting training data with these particular external datasets did not improve the performance. We note that further analysis of the relation between external and task data is needed to establish whether training data augmentation in general is a suitable strategy for this task.

Figure 1 shows the confusion matrices of the best performing ensemble models, both with the majority strategy, using the CASE dataset and using aggregated data.

A study of the errors of the different runs shows that the four sets were wrong simultaneously in 17.47% of the total number of test instances, three of them were wrong in 7.3% of the instances, two of them in 5.63% and only one of them in 8.32%. Grouping by label, we observe that 11.62% of the instances labelled as SUPPORT are misclassified by all models, 24.11% for those labelled as OPPOSE and 26.4% for NEUTRAL. Grouping by ensemble strategy, we notice that for the majority one, 22.02% of instances are misclassified by the two models, while it is 24.2% for the conservative one. For the majority voting, the error for SUPPORT is 11.62% of all instances labelled as SUPPORT, 24.11% for OPPOSE and 40.6% for NEUTRAL. In the case of the conservative strategy, SUPPORT is 23.02%, OPPOSE 24.11% and NEUTRAL 26.4%.

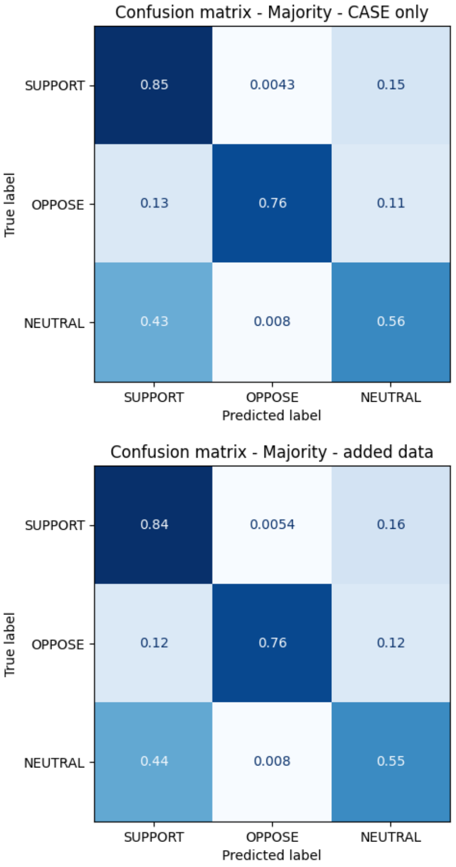

Figure 1: Confusion matrix for the best performing ensemble models using CASE and using added data.

Based on the errors per label, it can be seen that the conservative strategy, where predictions more often skew towards NEUTRAL, afforded a worse performance. Conservative ensembles were better at classifying NEUTRAL instances, but this was at the expense of the SUPPORT label. Since NEUTRAL instances are limited in the dataset, the use of the conservative strategy did not offer an advantage, whereas the majority more faithfully reproduced the expected distribution of labels in the dataset.

However, our RoBERTa baseline performed better than any of the ensemble strategies we submitted to the competition. In that light, we decided to conduct a post-competition examination with fewer restrictions to construct the ensemble system.

	Approach	F_1	Acc.
	Baseline	0.5450	0.6510
	Best model leaderboard	0.7495	0.7458
Indiv.	RoBERTa	0.7495	0.7458
	Llama 2	0.7366	0.7132
	MTL (best in training)	0.7402	0.7433
Ensemble	Conservative	0.7300	0.7004
	Mayority vote	**0.7529**	**0.7510**
	Oracle (upper bound)	0.8481	0.8451

Table 2: Comparison of F_1-scores and Accuracies for the best individual models (regardless of train data regime) and the ensembles of best models.

6 Post-competition Analysis

Our ensemble is based on the idea of combining the diversity given by a Transformer-based system (RoBERTa), a generative model (Llama 2) and a Multitask Learning (MTL) approach. Therefore, we just selected one configuration for each of the three approaches. Furthermore, we constrained ourselves to two options: whether all the systems use external datasets or none of them do.

After submission, we relaxed this constraint and performed a post-competition run selecting our best RoBERTa, Llama 2 and MTL models from the training stage, regardless of whether they use external data. In this case, as shown in Table 2, the majority vote ensemble achieves the best F_1 result, surpassing our RoBERTa based system that would have attained the highest position on the leaderboard for the stance detection subtask. If we look at accuracy, our majority vote ensemble surpasses the best model on the leaderboard.

The difference between both ensembles is due only to the use of a different configuration of the MTL model. This shows that the diversity introduced by the best in training MTL model is valuable to the ensemble.

7 Conclusions and Future Work

Our contribution focused on studying the influence of external data in the context of climate activism. We have done this through three different systems, whose combination into the proposed ensemble we present in this paper. The impact of external data on this particular subtask has been limited, only being effective in the case of the MTL system, which we theorize might be due to the different classification heads for each dataset, allowing them to keep the task-specific information of each task and maintaining the encoder with the general shared knowledge. In spite of this situation, we have gained some insight regarding our ensembles. Although our submitted runs do not improve the best individual result of the RoBERTa baseline, the post competition analysis reveals that an ensemble with our best models, regardless of the training set, would have achieved the first position in the competition.

As future work, a thorough study of the best combination of models, to find a higher divergence, is crucial. We have also determined that three systems might be insufficient for classification tasks with 3 classes, generating uncertainty in the test. To reduce this uncertainty, we plan on studying the effects of an ensemble with several models per approach, and of different voting strategies, such as a weighted voting schema, which could add a higher confidence level to the models and correct potential biases.

A central goal of our contribution, analyzing the effect of training with external data on this dataset, remains inconclusive. The proposed additional datasets did not always improve the results. We hypothesize that an analysis of the lexical and semantic distance between task data and external data could help to determine the suitability of the chosen collections. This analysis should potentially be extended over alternative external datasets in order to make an informed choice. A similar analysis of the particular instances of ClimaConvo in which each of the different models of the ensemble were successful – or failed – could contribute to better determine each model's strengths and clarify an optimal ensemble strategy.

As for individual models, another avenue to explore is studying the effect of other dimensions, such as pre-processing of the input data, as well as altering the threshold to assign a label to the instances. Although the conservative strategy did not have a high performance, the NEUTRAL tag still proves problematic. Optimizing the value for this threshold may improve the detection of this tag, thus enhancing the models. Additionally, an in-depth study of the effect of external data, and how each model performs for those tasks, would be necessary to determine why it is not as effective in the case of RoBERTa and Llama-2 and how we can improve it.

Limitations

An important drawback is the lack of regularization regarding external data usage in the constituent models of the ensemble presented in this paper. This situation limits the scope of the paper when addressing the value of additional data and requires a comprehensive analysis to determine its added value.

Another limitation relates to the high GPU requirements of some of our models. It is also relevant to note that some of the individual approaches achieve comparable results without such shortcomings. An additional study to determine if the usage of highly complex models for classification tasks may prove necessary.

Acknowledgements

This work was supported by the HAMiSoN project grant CHIST-ERA-21-OSNEM-002, AEI PCI2022-135026-2 (MCIN/AEI/10.13039/501100011033 and EU "NextGenerationEU"/PRTR).

References

Richard A. Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48.

Kaiping Chen, Amanda L. Molder, Zening Duan, Shelley Boulianne, Christopher Eckart, Prince Mallari, and Diyi Yang. 2023. How climate movement actors and news media frame climate change and strike: Evidence from analyzing twitter and news media discourse from 2018 to 2021. *The International Journal of Press/Politics*, 28(2):384–413.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Rosso Paolo, et al. 2020. SardiStance@ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 1–10. CEUR.

Jesus M. Fraile-Hernandez and Anselmo Peñas. 2024. HAMiSoN-Generative at ClimateActivism 2024: Stance Detection using generative large language models. *Preprint*.

Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E. Smaldino, Goran Muric, and Keith Burghardt. 2023. Auditing elon musk's impact on hate speech and bots. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1133–1137.

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. *CoRR*, abs/1907.11692.

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, Portorož, Slovenia. European Language Resources Association (ELRA).

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, San Diego, California. Association for Computational Linguistics.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308.

R. Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.

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. *Preprint*.

Raquel Rodriguez-Garcia and Roberto Centeno. 2024. HAMiSoN-MTL at ClimateActivism 2024: Detection of Hate Speech, Targets, and Stance using Multitask Learning. *Preprint*.

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*.

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)*.

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, 2023. *ArXiv*.

Aman Tyagi, Joshua Uyheng, and Kathleen M. Carley. 2020. Affective polarization in online climate change discourse on twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 443–447.

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). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.