HAMiSoN-MTL at ClimateActivism 2024: Detection of Hate Speech, Targets, and Stance using Multi-task Learning

Raquel Rodriguez-Garcia
NLP & IR Group
UNED, Spain
rrodriguez@lsi.uned.es

Roberto Centeno
NLP & IR Group
UNED, Spain
rcenteno@lsi.uned.es

Abstract

The automatic identification of hate speech constitutes an important task, playing a relevant role towards inclusivity. In these terms, the shared task on Climate Activism Stance and Hate Event Detection at CASE 2024 proposes the analysis of Twitter messages related to climate change activism for three subtasks. Subtasks A and C aim at detecting hate speech and establishing the stance of the tweet, respectively, while subtask B seeks to determine the target of the hate speech. In this paper, we describe our approach to the given subtasks. Our systems leverage transformer-based multi-task learning. Additionally, since the dataset contains a low number of tweets, we have studied the effect of adding external data to increase the learning of the model. With our approach we achieve the fourth position on subtask C on the final leaderboard, with minimal difference from the first position, showcasing the strength of multi-task learning.

1 Introduction

The shared task on Climate Activism Stance and Hate Event Detection at CASE 2024 (Thapa et al., 2024) focuses on climate change discussions on Twitter. As of late, climate change is experiencing an increase in political polarization (Falkenberg et al., 2022), and these trends have revealed connections to a higher power controlling the public's discourse (Farrell, 2016). This situation highlights the importance of an in-depth study of the issue and the many challenges it still poses, from the data collection to the added difficulty of multilingual approaches (Parihar et al., 2021). This task, which studies the content of tweets in relation to hate speech and other essential characteristics, such as the target of the message, can serve to provide more insights regarding how these messages are transmitted and their common features.

Recent competitions have been held for the detection of hate speech or offensive language (Lai et al., 2023) as well as the target of the message (Bhandari et al., 2023; Zampieri et al., 2019b), focusing on issues such as multilingual Twitter data or multimodal content. State-of-the-art results are obtained through the use of transformer-based approaches, that are capable of employing the entire context of the data. Additional contextual knowledge, such as social information or newspaper articles, has also shown its effectiveness to improve a system's performance (Nagar et al., 2023; Pérez et al., 2023). Similarly, stance detection has been a traditional research topic for shared tasks (Cignarella et al., 2020; Davydova and Tutubalina, 2022), where transformer-based approaches, along with data augmentation, tend to outperform other methods.

In our approach to this task, we leverage the potential of multi-task learning (MTL) with a pre-trained transformer model for the subtasks. MTL, as originally presented by Caruana (1993), is able to extract information from one task to boost the performance of another, without the necessity of transferring the knowledge attained and the complications it poses with the differences in tasks or annotations. It also reduces the risk of overfitting (Baxter, 1997) due to the shared representation it generates for all the tasks.

In our systems, we experiment with added datasets, to fully exploit the capabilities of MTL. We explore the effects of additional data for each of the tasks, with different levels of relatedness. To fully study that effect, we also fine-tune our systems without external data, other than the three subtasks. We aim to discover what works best in this situation, where we have three highly related subtasks, but there is a lack of data, especially for subtask B.

The paper is structured as follows: in section 2 we briefly discuss the characteristics of the shared task, as well as the dataset provided. In section 3 we describe our approach by leveraging MTL and

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 89–95
March 22, 2024 ©2024 Association for Computational Linguistics

Subtask	Class	train	dev	test
A	Non Hate Speech	6385	1371	1374
	Hate Speech	899	190	188
B	Individual	563	120	121
	Organization	105	23	23
	Community	31	7	6
C	Support	4328	897	921
	Oppose	2256	153	141
	Neutral	700	511	500

Table 1: Annotation statistics of the dataset for each subtask and set: train, dev and test.

external data sources. In section 4 we include the results of our systems and discuss the approaches. Finally, we highlight the conclusions in section 5.

2 Dataset & Task

The dataset for the shared task on Climate Activism Stance and Hate Event Detection, introduced in Shiwakoti et al. (2024), contains a total of 10,407 tweets, only including the textual content. These instances were collected using hashtags linked to climate change and related activism and only selecting English tweets. Finally, they were manually annotated for different tasks. We describe below each of the subtasks that are part of the shared task. The tweet distribution for each subtask is shown in Table 1.

2.1 Subtask A: Hate Speech Detection

Subtask A is aimed at determining whether a tweet is considered hate speech or not. The tweets are annotated for this binary classification task with two labels: *Hate Speech* and *No Hate Speech*.

2.2 Subtask B: Target Detection

The objective of this subtask is to establish the target of the hate speech. The annotation for this multi-class classification task is given by three classes: *Individual*, *Organization* or *Community*. In these tweets there is hate speech, therefore, only a part of the tweets in the full dataset are annotated with the target.

2.3 Subtask C: Stance Detection

The goal of this last subtask is to establish the stance of each tweet. This is also a multi-class classification task with three possible classes: *Support*, *Oppose* or *Neutral*. These are the same tweets used for subtask A.

3 Methodology

For our experiments, we use the same pre-trained transformer model throughout the different combinations for comparability purposes. The selection of the model is influenced by two main factors: generalization and robustness. Models trained on domain-specific data or from select data sources, such as Twitter, would not be ideal for our study, since we incorporate other corpus not Twitter nor climate related. Additionally, we want to ensure the selected model provides robustness in terms of textual classification tasks. These considerations justified our selection of the RoBERTa (Liu et al., 2019) pretrained model we used.

The architecture of the MTL system corresponds with a hard parameter sharing approach: for each task we make use of one classification head and a RoBERTa shared encoder for all of them. Since the data sources are different in most cases, each input instance only corresponds with one classification task. The model uses size-proportional sampling, in regard to each of the datasets for the classification tasks, when selecting the next instance during training, with a fixed batch size of 32.

As we previously introduced, we are using external data for the task. We briefly describe them below.

- Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a). This dataset, composed of Twitter data, was used in the SemEval 2019 Task 6, OffensEval (Zampieri et al., 2019b). It has three tasks: offensive language identification (*Offensive* or *Not Offensive*), categorization of offense types (*Targeted* or *Untargeted*) and offense target identification (*Individual*, *Group* or *Others*). Due to the similarity between the offense and target identification tasks to subtask A and B, we select these OLID tasks for our training. We combine the train and test partitions into one dataset for the training of our system, generating a total of 14,100 and 4,089 tweets for the offense and target tasks, respectively.
- The stance dataset presented in Mohammad et al. (2016a), which was used in SemEval-2016 Task 6 (Mohammad et al., 2016b). For easier reference throughout the paper, we will refer to it as StancEval. This dataset is divided into different sections depending on the topic of the tweet. These include abortion, Hillary Clinton, atheism,

climate and feminism, for a total of 4,163 tweets. The classification of the tweets considers three classes: *Against*, *Favor* or *None*. The train and test data are combined for our training.

- COP27 data. This source of data is composed of unannotated tweets gathered during COP27, using related hashtags. Given that the tweets had no relevant annotation, we decided to assign a simple label for the ease of use as a classification task. We created a binary task to determine the presence or absence of a retweet. Although the task is unrelated and the annotation might be irrelevant, the tweets are related, and it might provide additional context to the system. To establish if unannotated data could be useful, we select a total of 45,000 random tweets. We aim to determine if having more available data can compensate for the weak annotation or lower relatedness to the task.
- The Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018). This dataset consists of a textual premise and a hypothesis, and the class indicates if there is *Entailment*, *Contradiction* or a *Neutral* relationship between them. Contrary to previous datasets, this one is unrelated to the task. To make it comparable, we select a class-balanced sample of 12,000 instances.

These datasets are combined into the models displayed in Table 2. Below, we explain each of them.

- **BASE**. For this run, we only consider the base data for this CASE task, with one model for the three subtasks.
- **BASE StancEval climate**. Since StancEval contains information not related to climate change, we only select the climate topic, in addition to the base subtasks.
- **BASE StancEval full**. For this run, we include the whole StancEval dataset with the base subtasks.
- **BASE OLID**. This run includes the offense and target identification subtasks from OLID.
- **BASE OLID, StancEval**. For this run, we use the full OLID and StancEval datasets and the three subtasks.
- **BASE MultiNLI**. For this model, we use the three subtasks and the MultiNLI task.
- **BASE COP27**. This run adds the unrelated annotation from the COP tweets to the three subtasks.
- Only one base task and the closest task from another dataset. For this run, we select only one of the individual subtasks from the task and run an MTL model with another similar task. For subtask A (**Hate Only**) and B (**Target Only**), we use the OLID offense and the target identification, respectively. For subtask C (**Stance Only**) we use the full StancEval dataset.
- Best model configuration retrained on all data (**Best model**). The best model obtained during the evaluation, without accounting for the final test results, is run with the full training data.

Regarding the preprocessing of the textual input, only the Twitter data is altered. Since it includes hashtags and user mentions that the transformer might not be able to represent, we need to consider a previous step for normalization. All the mentions and URLs have been removed from the text. For the hashtags, we have followed a different approach by splitting the text into words using wordninja (Keredson, 2019), since hashtags are usually a concatenation of words that might provide additional insight into the user's opinion. In the case of the MultiNLI dataset, the premise and the hypothesis are combined into an input with a separator in-between the texts for the model.

For our experiments, we explore the combinations of an initial set of parameters shown in Table 3. Although more combinations were initially tested, we discarded them due to low results. For the final submissions, we select the parameter combination with the highest F1 on our evaluation data, for each subtask, and submit the results for all the combinations outlined above. We aim to use a comparable configuration to better analyze the results of the different combinations described.

Since the dev labels were not available when we first trained our systems, we created our class balanced partition of 70-30 for the training and evaluation of the subtasks (except for subtask B, which had fewer instances, so we decided on 80-20). After they were made public, we also uploaded our systems using the dev partition for evaluation and the train set for training. We report all the results in the next section for a more in-depth analysis. Additionally, for the best model retrained, in our first partition we use all the training data, while in the second we use the training and dev data combined.

Run	CASE			StancEval		OLID		MultiNLI	COP27
	A	B	C	climate topic	all topics	offense	target		
BASE	✓	✓	✓						
BASE StancEval climate	✓	✓	✓	✓					
BASE StancEval full	✓	✓	✓		✓				
BASE OLID	✓	✓	✓			✓	✓		
BASE OLID, StancEval	✓	✓	✓		✓	✓	✓		
BASE MultiNLI	✓	✓	✓					✓	
BASE COP27	✓	✓	✓						✓
Hate Only	✓					✓			
Target Only		✓					✓		
Stance Only			✓		✓				

Table 2: Different models tested and their data sources.

Parameter	Values
Epochs	3 and 4
Learning rate (LR)	2e-5 to 5e-5, step 1e-5
Weight decay	1e-3
Epochs	3 and 4
Learning rate (LR)	3e-5 and 4e-5
Weight decay	1e-2 and 1e-4

Table 3: Ranges of parameters used for training.

Task	Partition	LR	Epochs	Weight decay
A	70-30	3e-5	4	0.001
	train-dev	4e-5	4	0.0001
B	80-20	3e-5	4	0.0001
	train-dev	3e-5	3	0.0001
C	70-30	2e-5	3	0.001
	train-dev	4e-5	4	0.001

Table 4: Final parameter configuration for the submitted runs, for each task and partition.

4 Results & Discussion

The F1 results for the final configuration of the parameters uploaded for each subtask and partition is detailed in Table 4, based on the results of the evaluation (the 30% partition or the dev set), which are gathered in Table 5. For the A and C subtasks, regardless of the partition, values are very similar for most runs. There are slight differences between the partitions, which could be caused by differences between the tweets in the sets. Additional data does not appear to have a pronounced effect, although it achieves the best results. In subtask C for the dev partition, the COP27 run seems ineffective, which might indicate the difference in the data. In subtask B there is a higher variance between results. We can better appreciate the improvement of external datasets, especially with the most related ones, maybe due to the low amount of data. In this case, unrelated data does not have a positive effect.

The results for the F1 metric on the test set for each of the runs described above, based on the partitions, are gathered in the Table 6. The baselines included are the ones reported in Shiwakoti et al. (2024) and we can observe how our systems significantly outperform them. For subtask A, most of the results are similar, which might indicate the models are already reaching their plateau. We can also appreciate that less relevant data (MultiNLI or COP27) achieves relatively good results, which might indicate additional data is not necessary, or it hinders performance, especially considering that our best result is achieved with only the original data, attaining the sixth position in the leaderboard.

In subtask B there is a much higher difference between the results. The low amount of data, particularly compared to the other tasks the model was trained with, might have caused an imbalance when the model was learning for this task. Adjusting the size of the datasets, or augmenting the data, may have a positive impact. It is also interesting to note that the best result is achieved when training with 80% of the training set and the most similar task. Seemingly, adding highly related data has the best impact, securing the eighth position in the ranking.

In subtask C we notice that most results are similar, although COP achieves the lowest in one run. We can observe again that additional data does not have a very high impact, but it achieves the highest result with the fourth position in the leaderboard and minimal difference to the best system.

In terms of error analysis for the subtasks, we have noticed some tendencies. For subtask A, in over half of the runs, *Hate Speech* is correctly detected for a total of 98% of the class instances. Meanwhile, all runs predict the wrong class for *Non Hate Speech* in 10% of the instances for that class. Even though *Non Hate Speech* is the majority class, the system struggles to differentiate it. For subtask B we observe a similar effect, with over half of the runs being able to detect the *Individual* and *Organization* for over 90% of those instances.

Approach	Task A		Task B		Task C	
	part	dev	part	dev	part	dev
BASE	0.8666	0.8609	0.5227	0.6742	0.7080	0.6908
BASE StancEval climate	0.8682	0.8643	0.6665	0.5365	0.7187	0.6989
BASE StancEval full	0.8597	0.8483	0.6908	0.5326	0.7100	0.6824
BASE OLID	**0.8781**	**0.8738**	**0.8711**	0.8304	0.7083	**0.7137**
BASE OLID, StancEval	0.8739	0.8637	0.7197	0.8136	0.7073	0.6973
BASE MultiNLI	0.8566	0.8587	0.5882	0.5458	0.7162	0.6983
BASE COP27	0.8485	0.8202	0.5315	0.5327	0.7130	0.5102
Hate Only	0.8572	0.8675				
Target Only			0.7189	**0.8699**		
Stance Only					**0.7213**	0.6986

Table 5: Results for the subtasks, for the evaluation set (the 20-30% partition or the dev set).

Approach	Task A		Task B		Task C	
	part	full	part	full	part	full
Baseline	0.708		0.554		0.545	
Best Systems	0.9144		0.7858		0.7483	
BASE	0.8713	**0.8840**	0.5505	0.6668	0.7220	0.7274
BASE StancEval climate	0.8638	0.8788	0.6052	0.5752	0.7263	0.7212
BASE StancEval full	0.8757	0.8706	0.6280	0.5565	**0.7351**	0.7322
BASE OLID	0.8757	0.8731	0.7124	0.7046	0.7218	**0.7402**
BASE OLID, StancEval	0.8725	0.8806	0.6828	**0.7206**	0.7156	0.7324
BASE MultiNLI	0.8632	0.8656	0.5431	0.5345	0.7319	0.7263
BASE COP27	0.8672	0.8461	0.6259	0.5496	0.7298	0.5394
Hate Only	0.8609	0.8574				
Target Only			**0.7329**	0.6640		
Stance Only					0.7309	0.7214
Best model	**0.8794**	0.8774	0.7111	0.6375	0.7240	0.7320

Table 6: Results for the subtasks, for the test set (training on the 80-70% partition or the train set). The best model retrained refers to the model from Table 5 with the highest score.

In this case, we notice the system errs while identifying the *Community*, although that could be due to being the minority class. Finally, over half the runs for subtask C tend to coincide for the *Support* and *Oppose* classes with 88% and 75% of accuracy respectively, although it decreases to 50% for *Neutral*. Our runs tend to predict *Support* when the class is *Neutral*, which could be due to noisy data or some level of ambiguity in the texts.

In summary, it appears that external data has achieved the best result in subtasks B and C. Even when the dataset was not as related to the subtask, it still appeared to add some additional knowledge. There is a high difference between the evaluation and the test results for subtask B, which could indicate some problems already mentioned for the data or a low sampling for the MTL models. Regarding subtask A, since most of the results were very similar, the differences between the runs might be more related to randomness rather than the ineffectiveness of the additional data.

5 Conclusion

Hate speech is a growing cause of concern on social media, and it is still on the rise, spreading polarization to seemingly uncontroversial new topics, such as climate change. With our approach to this task, we propose to leverage other existing datasets through transformer-based MTL. Our models present a robust approach to address data scarcity, especially for the target detection subtask, without the need to adapt annotations or merge unrelated data, while creating models with a higher capacity to generalize. Our findings reveal that external data that is highly related to the task has an overall positive effect, while the lower the relatedness, the worse results we achieve.

As a result from our experiments, our models have shown that the most promising performances are achieved when external data is used to improve one of the tasks. As future work, we plan on having a more balanced dataset for target identification, as well as experimenting with other pre-trained or already fine-tuned models for specific tasks that might provide additional context, such as sentiment analysis. Additionally, we want to study the effect that each external dataset had on the models' predictions and their contributions to the results, which might provide insights into how to further improve this approach.

Limitations

The high variance in results from validation to test in subtask B indicates the presence of overfitting, possibly reducing the ability of the model to generalize in that task. Adjusting the sizes of the datasets, through augmentation or oversampling, or tuning the sample sizes would be necessary to address this issue.

Since the goal was to optimize each of the subtasks for the shared task, models were not evaluated for each of the auxiliary tasks and datasets included. Additional testing would be necessary to create a more robust approach and to determine if the MTL system improves other tasks' performances, although that might impact the effectiveness of the models for this shared task, therefore, the tradeoff should be considered.

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

Jonathan Baxter. 1997. A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling. *Machine learning*, 28:7–39.

Aashish Bhandari, Siddhant B. Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. CrisisHateMM: Multimodal Analysis of Directed and Undirected Hate Speech in Text-Embedded Images from Russia-Ukraine Conflict. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1994–2003.

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

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 *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, page 177–186. Accademia University Press.

Vera Davydova and Elena Tutubalina. 2022. SMM4H 2022 task 2: Dataset for stance and premise detection in tweets about health mandates related to COVID-19. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 216–220, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, and Andrea Baronchelli. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121.

Justin Farrell. 2016. Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences*, 113(1):92–97.

Keredson. 2019. Wordninja Python Package.

Mirko Lai, Fabio Celli, Alan Ramponi, Sara Tonelli, Cristina Bosco, and Viviana Patti. 2023. HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR, Parma, Italy*, volume 3473.

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*, arXiv: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.

Seema Nagar, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2023. Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining*, 13(1):47.

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.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the Impact of Contextual Information in Hate Speech Detection. *IEEE Access*, 11:30575–30590.

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, Hari 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)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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.