HAMiSoN-baselines at ClimateActivism 2024: A Study on the Use of External Data for Hate Speech and Stance Detection

Julio Reyes-Montesinos and **Álvaro Rodrigo**
NLP & IR Group
UNED, Spain
{jreyes, alvarory}@lsi.uned.es

Abstract

The CASE@EACL2024 Shared Task addresses Climate Activism online through three subtasks that focus on hate speech detection (Subtask A), hate speech target classification (Subtask B), and stance detection (Subtask C) respectively. Our contribution examines the effect of fine-tuning on external data for each of these subtasks. For the two subtasks that focus on hate speech, we augment the training data with the OLID (Zampieri et al., 2019a) dataset, whereas for the stance subtask we harness the SemEval-2016 Stance dataset (Mohammad et al., 2016b). We fine-tune RoBERTa and DeBERTa models for each of the subtasks, with and without external training data. For the hate speech detection and stance detection subtasks, our RoBERTa models came up third and first on the leaderboard, respectively. While the use of external data was not relevant on those tasks, we found that it greatly improved the performance on the hate speech target categorization.

1 Introduction

In recent years, the escalating global awareness of the imminent climate crisis has not only prompted an upsurge in climate activism but has also given rise to a new wave of advocacy strategies, often marked by actions not devoid of controversy. While the urgency of addressing climate change has fostered a sense of shared responsibility in society, some of the actions of climate activists have also sparked debates regarding the boundaries of acceptable dissent. When translated to the online sphere, where climate activists looking to disseminate their messages and mobilize supporters encounter both climate deniers and corporate PR, these conversations become ever more heated, often precluding sensible debate. Our research aspires to contribute to a deeper understanding of the digital discourse surrounding climate activism and facilitate the creation of tools that can foster healthier online conversations while respecting the fundamental right to dissent in an age of environmental urgency.

This paper delves into the intricate landscape of online climate activism, with a focus on the automated detection of hate speech in this context. Specifically, our contribution looks at the effect of fine-tuning transformers on two external datasets selected for their relatedness to the tasks at hand, besides the data proposed by the task itself. For the subtasks focusing on hate speech detection and the categorization of its target, we augmented the training data with the OLID (Zampieri et al., 2019a) dataset. In turn, for the stance detection subtask we employed the section related to climate change of the SemEval-2016 Stance dataset (Mohammad et al., 2016b).

The rest of this paper describes the data provided by the task (section 2) as well as the external data (section 3) we chose to augment it. Next, we detail the system development process (section 4) and discuss the results (section 5. We finish with a brief Conclusion (section 6.

2 Dataset and Task

The ClimaConvo dataset Shiwakoti et al. (2024) exposes a cross-section of the public discourse around climate change on social media. It comprises 15,309 tweets collected around a series of hashtags related to climate activism over a one-year period. The dataset contains annotations in six layers: relevance, stance, the presence of hate speech; if present, whether it is directed; when directed, the type of target; and the presence of humor.

The shared task at hand, CASE@EACL2024 (Thapa et al., 2024), comprises three subtasks based on two subsets of ClimaConvo, corresponding to 10,407 tweets. These subsets have been split in train, validation and test sets by the authors. Table 1 describes the subsets, splits, and the balance of labels in them. Each of the tasks relates to one of the annotation layers in ClimaConvo, as follows:

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 156–160
March 22, 2024 ©2024 Association for Computational Linguistics

2.1 Subtask A

The first subtask involved the detection of hate speech in tweets. It therefore contains all tweets labeled RELEVANT in ClimaConvo, which can in turn be labeled as containing HATE SPEECH or containing NO HATE SPEECH.

2.2 Subtask B

For this subtask, participants were asked to categorize the target of hate speech in tweets, resulting in a multi-class classification task with the labels INDIVIDUAL, ORGANIZATION and COMMUNITY. The subtask is based on the subset of tweets in ClimaConvo where hate speech is labeled as RELEVANT, that is, a smaller subset of the one introduced for the previous task, this time adding up to 999 tweets.

2.3 Subtask C

The stance subtask is based on the same subset of tweets as subtask A, i.e. RELEVANT tweets. The train, validation and test splits also remain constant. However, this subtasks asks participants to determine whether tweets SUPPORT or OPPOSE Climate Activism, or have NEUTRAL position towards it.

3 External Data

We sourced the additional training data for our experiments from two datasets external to the task: the OLID and the SemEval-2016 Task 6 datasets.

3.1 OLID

As external data related to hate speech, we consider the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019b) presented at SemEval-2019 (Zampieri et al., 2019c). OLID was Compiled with the goal of tackling the problem of offensive posts in social media as a whole, OLID consists of 14,100 tweets annotated in three layers: the presence of offensive language; if present, its categorization (as Targeted or Untargeted); and if targeted, the identification of this target (an Individual, a Group or Other type of entity). We manually compared a sample of tweets to match these labels to their Individual, Organization and Community counterparts in ClimaConvo.

For Subtask A, we use the full OLID dataset (since all tweets are annotated for presence of offensive speech). For Subtask B, we use the subset of 4,089 tweets identified as targeted, and therefore annotated for target type. Although the authors define train and test splits, we merge both splits as additional train data.

3.2 SemEval-2016 Task 6

For the stance detection subtask (Subtask B), we harness the Stance Dataset Mohammad et al. (2016a) presented at the SemEval-2016 Task 6 (Mohammad et al., 2016c). This dataset consists of a total of 4,870 tweets labeled with the stance they express about a certain target topic: abortion, climate, Hillary Clinton, feminism, atheism, and Donald Trump. For our purpose of training data augmentation, we use only the portion related to climate change, which totals 564 tweets. The labels (Favor, Against or Neither) are analogous to the ones in ClimaConvo.

4 Methodology

The present contribution has the goal of establishing state-of-the-art transformer baselines for the three subtasks, and then examine the influence of additional training data on each subtask. To this end, we developed systems based on the RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) transformers.

Both RoBERTa and DeBERTa improve upon BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) by introducing different training objectives: RoBERTa uses dynamic masking (where different tokens are masked every time the same sequence is fed to the model) and eliminates the next-sentence prediction training objective of BERT. DeBERTa adds a disentangled attention mechanism (where each word is represented using two vectors that encode its content and relative position) and enhanced masked decoding (where absolute word positions are added back). The version we use, DeBERTa-v3 (He et al., 2021), replaces the masked language model pre-training task with replace token detection task (RTD), further improving the models capacity to capture long-range dependencies over RoBERTa. On the other hand, it is key to note that RoBERTa has been pre-trained on double the amount of data.

Common to both of these transformer architectures is the notion that they can be fine-tuned at a low computational cost while still exceeding at a number of diverse Natural Language Understanding tasks. In the following subsections we provide technical details of how the proposed models were fine-tuned on the reference datasets:

subtask	A: hate speech		C: stance			
label	NO HATE SPEECH	HATE SPEECH	SUPPORT	OPPOSE	NEUTRAL	Total/split
train	6385	899	4328	2256	700	7284
validation	1371	190	897	511	153	1561
test	1374	188	921	500	141	1562
Total/label	9130	1277	6146	3267	994	10407

subtask	B: hate speech target			
label	INDIVIDUAL	ORGANIZATION	COMMUNITY	Total/split
train	563	105	31	699
validation	120	23	7	150
test	121	23	6	150
Total/label	804	151	44	999

Table 1: Per split label distribution in tweets assigned to each subtask.

4.1 Dataset pre-processing

Before feeding the data to the models, we followed a common text pre-processing pipeline for tweets, on both the task and the external data, consisting of the following actions:

- Replacement of URLs by the special tokens [URL_TWITTER] and [URL_OTHER].
- Replacement of username mentions by the generic token @USER.
- Splitting of hashtags into individual words. To accomplish this endeavour we have utilized the Word Ninja[1] library, which uses a probabilistic division of concatenated words, based on the frequencies of unigrams in the English Wikipedia.

4.2 Fine-tuning configuration

We first fine-tuned off-the-shelf RoBERTa-base[2] and DeBERTa-v3-base[3] transformers with text classification heads for each of the subtasks using only the data proposed in the shared task. We then fine-tuned a second set of RoBERTa and DeBERTa models including the proposed additional training data for each subtask.

Some of the models' hyper-paramenters have been determined experimentally: All models have been fine-tuned for 3 epochs. Tweets are administered in a random order, and when using external data, these are lumped together with the subtask's original data. The batch size is 8 for RoBERTa, but 4 for DeBERTa due to memory constraints. The learning rates are 2×10^{-5} for RoBERTa and 1×10^{-5} for DeBERTa.

All learning rates are scheduled to first linearly increase from 0 to the aformentioned rates during an initial pediod of 100 training steps, and then decrease linearly for the rest of trainign steps. The chosen optimizer in all cases is AdamW.

During development, models were fine-tuned on the proposed train split only. The models submitted in the test phase, however, have been fine-tuned on both the train and the validation splits proposed by each subtask (as well as the proposed external data when applicable).

4.3 Submitted runs

Summing up, for each of the three subtasks we submitted four runs:

1. RoBERTa-base fine-tuned on subtask's data.
2. DeBERTa-v3-base on subtask's data.
3. RoBERTa-base on subtask's + additional data.
4. DeBERTa-v3-base fine-tuned on subtask's + additional data.

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

[2] https://huggingface.co/FacebookAI/roberta-base

[3] https://huggingface.co/microsoft/deberta-v3-base

5 Results and Discussion

Results on subtasks A (hate speech detection) and C (stance detection) follow a similar pattern: our best results are achieved by the RoBERTa models fine-tuned on subtask data only. As seen on table 2, models fine-tuned on external data perform worse that their counterparts trained on subtask data only, but more so the RoBERTa's. DeBERTa models perform similarly regardless of whether we fine-tuned them on additional data, while the divergence is bigger for RoBERTa's.

On these subtasks, our models come far above all of the baselines provided by the organizers. On subtask A, our models come close below the best in the leaderboard. On subtask C, our simple RoBERTa comes atop the leaderboard. We note that these results are also far superior to the RoBERTa baseline provided by the organizers — we attribute this difference to our more thorough pre-processing and the difference in hyper-parameters. We also note, however, that the organizer's baseline that is already fine-tuned on climate-related text (ClimateBERT) obtains better results than other baselines on these two subtasks.

The pattern of results on subtask B (hate speech target categorization) is different: here the impact of external data is notably positive in the results. The RoBERTa fine-tuned on additional data is our best model on this subtask, whereas the models trained on subtask data only do not improve on the organizer's baselines. We attribute this difference to the size of the subset of tweets designated for this task. The much larger size of the chosen additional dataset (4,089 vs. 999 tweets) is more attuned to what transformer models such as RoBERTa and DeBERTa expect.

Finally, we consider that RoBERTa models perform better than more advance DeBERTa models on this task due to contextual knowledge being more important than the ability to capture long-range dependencies when dealing with tweet data, whose instances are short in nature.

6 Conclusions and future work

This paper introduced carefully adjusted transformer baselines for the hate speech detection, hate speech target categorization, and stance detection in tweets subtasks proposed at CASE@EACL2024. Based on off-the-shelf models, we have conducted a study of the effects of related external train data, with mixed results. We consider that further analysis of the results is needed before discarding the use of external data for this task. In particular, we would like to study the lexical and semantic distance between the ClimaConvo dataset proposed by the task and the ones chosen as additional train data, aiming to extend this analysis to other potential external datasets.

This research contributes to the ongoing efforts to foster healthy online conversations surrounding climate change activism. As the field of natural language processing continues to advance, our systems serve as a foundation for future developments in hate speech and stance detection in the context of critical issues like climate change.

	F_1 score by subtask		
Model	**A**	**B**	**C**
Best model on leaderboard	0.9144	0.7858	0.7483
Task's Baselines:			
BERT	**0.708**	**0.554**	0.466
DistillBERT	0.664	0.550	0.527
RoBERTa	0.662	0.501	0.542
ClimateBERT	0.704	0.549	**0.545**
RoBERTa	**0.8886**	0.5518	**0.7495**
DeBERTa	0.8751	0.5493	0.7408
RoBERTa ext.data	0.8682	**0.7017**	0.7406
DeBERTa ext.data	0.8713	0.6588	0.7392

Table 2: F_1 scores achieved by our submitted runs on each subtask compared to the baselines provided by the organizers and those achieved by the best participating systems. In bold, best baseline and best of our systems.

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

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

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.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016c. 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.

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)*.

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. 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. 2019c. 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.