JRC at ClimateActivism 2024: Lexicon-based Detection of Hate Speech

Hristo Tanev
European Commission, Joint Research Centre,
via Enrico Fermi 2749,
Ispra 21020, Italy
hristo.tanev@ec.europa.eu

Abstract

In this paper we describe the participation of the JRC team in the Sub-task A: "Hate Speech Detection" in the Shared task *Stance and Hate Event Detection in Tweets Related to Climate Activism* at the CASE 2024 workshop. Our system is purely lexicon (keyword) based and does not use any statistical classifier. The system ranked 18 out of 22 participants with F1 of 0.83, only one point below a system, based on LLM. Our system also obtained one of the highest achieved precision scores among all participating algorithms.

1 Introduction

In this paper we report on the participation of the Joint Research Centre team at Subtask A: *Hate speech detection* in the shared task *Stance and Hate Event Detection in Tweets Related to Climate Activism* at CASE 2024 (Thapa et al., 2024) using a simple lexicon - based hate speech detection approach.

Over the past few years, the convergence of NLP and sociopolitical discourse has led to the development of diverse technologies such as hate speech detection, sentiment analysis, and other opinion detection technologies. At the same time, climate activism has taken a momentum on the social Web and has captured the attention of NLP researcher community working in these areas (Shiwakoti et al., 2024). As the public discussions in this topic proliferated, the escalation of hate speech started to raise concerns among users.

Within the climate change discourse, hate speech manifests as a concerning trend, often taking aim at specific entities such as climate activists, influential environmental and political organizations like Greenpeace, and even entire governmental bodies responsible for environmental policies. The targeting extends beyond institutions to include environmental initiatives like FridaysForFuture (Niininen and Baumeister, 2022), amplifying the scope of the issue.

Adding another layer to this complex scenario, there is a noteworthy phenomenon involving individuals who pretend allegiance to the climate activist cause. These people employ hate speech in a troll like manner as a weapon in defending their version of climate advocacy.

This dual nature of hate speech within the climate change discourse unveils the intricate interplay between genuine concerns, political discontent, and the broader socio-political landscape. This highlights the need for nuanced approaches in addressing hate speech, considering its diverse sources and motivations within the context of environmental activism.

In this picture, automatic hate speech detection is becoming important, keeping "clean" the space of the social platforms and preventing online users from exposure to extreme content and disinformation. On the other hand, hate speech shows also increase of the discontent and frustration towards certain topics and public personalities. It serves as an indicator of the significance of these issues and people and their public perception; it also plays a crucial role as a marker for a negative bias in the social discourse. In fact, in USA certain hate speech acts are given constitutional protection (Rosenfeld, 2002) under the laws defending the freedom of speech.

The purpose of our experiment was to put in comparison a keyword based system with the other shared task participants, which were expected to predominantly exploit machine learning methods. As the simplicity of our method suggests, our system achieved score only little above the average system performance, and ranked 18th out of 22 systems, with F1 score of 0.83. Our score was 0.03 lower than the system in the middle of the ranking; our method scored F1 lower by 0.07 from the top ranked system. The experiment proved that

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 85–88
March 22, 2024 ©2024 Association for Computational Linguistics

Why are powerful men so scared of Greta Thunberg? The FridaysForFuture movement and the idea that we'd all have the gall to conduct a ClimateStrike every Friday frightens and infuriates plutocrats.
How Billionaires with Greta Thunberg uproot the system, important thread 2 read:
Mitigate or die! Adaptation, even successful, to today's accelerating climate crisis is a deadly delusion for complacent inaction. Possible survival = immediate emissions decline!
the struggle continues greed capitalism and stupidity r the main reasons the planet is dying
For some third-rate TV presenters, attacking Greta Thunberg is the only way to get back into conversation again. In 50 years, no one will know who Brendan O'Neill was, but Greta Thunberg will still be known.
First we destroy nature The rich keep getting richer The poor are increasing in numbers Measly check in the mail When there's still hell to pay Bills and pills Then we destroy ourselves Push back Despite all this

Table 1: Example of hate speech detected by our system

lexicon based detection is less accurate than statistical methods, still not very far behind: we have obtained a score only 0.01 lower than the preceding in the ranking system, which used a large language model; moreover, our precision was the among the highest ones.

2 Related work

Hate speech is a topic of debate among lawmakers (Rosenfeld, 2002) and NLP experts (Jahan and Oussalah, 2023), (Parihar et al., 2021). Automatic hate speech detection has been predominantly approached as a binary text classification, using machine learning (Fortuna and Nunes, 2018); multilingual dimension has also been explored in previous works and shared tasks (Siino et al., 2021)

Lexicon-based hate speech analysis has also been addressed in previous works (MacAvaney et al., 2019), (Gitari et al., 2015). According to (MacAvaney et al., 2019), keyword-based approaches offer elevated precision but suffer from insufficient recall due to challenges in resolving word sense ambiguity and handling figurative language. Essentially, systems relying on keywords may overlook hateful content that doesn't employ explicit hate terms. In contrast, (Gitari et al., 2015) presents a lexicon-based approach that contradicts this assertion by demonstrating reasonably high levels of both precision and recall.

Hate speech detection is also strongly related to sentiment analysis and opinion mining, where lexicon-based approaches are still used: a comprehensive study of these techniques is presented in (Bonta et al., 2019).

3 Dataset and Task

The purpose of the Shared task on Detecting Hate Speech During Climate Activism was identification of tweets discussing the climate change topic and containing hate speech. The tweets have been retrieved by a team of researchers from Delhi Tech University, Virginia Tech, and James Cook University, Australia. The retrieval and annotation are described in (Shiwakoti et al., 2024). The data collection process aimed at tweets posted between January 1, 2022, and December 30, 2022. The selection criteria involved hashtags such as #climatecrisis, #climatechange, #ClimateEmergency, #ClimateTalk, #globalwarming, as well as activist-oriented hashtags like #fridaysforfuture, #actonclimate, #climatestrike, #extinctionrebellion, #ClimateAlliance, #climatejustice, #climateaction, etc. Only tweets composed in the English language have been considered by the data collection team. In this way above 15,000 tweets have been collected, which were subsequently annotated for presence of hate speech, relevance to the climate change discourse, stance, the direction of hate speech, targets of hate speech, and humor. For our shared task, only three aspects were considered from this annotation: hate speech, target of the hate speech (who or what is targeted) and stance (does the tweet support, oppose or is neutral). Given we participated in subtask A: Hate speech detection, only the hate speech annotation (1 - presence of hate speech, 0 - absence) was considered.

4 Methodology

In our approach we have used the Liu and Hu Lexicon (Ding et al., 2008), which is ranked as a high performing sentiment analysis lexicon by several studies: It was evaluated on Twitter data with information about people and other entities (Al-Shabi, 2020), as well as on product reviews (Khoo and Johnkhan, 2018). In both cases this lexicon has delivered very competitive results, with respect to other repositories of sentiment keywords. Considering our shared task on climate activism, the above mentioned Twitter based evaluation showed that the lexicon was relevant for the task.

The Hu and Liu lexicon has been created by two researchers from the Department of Computer Science of the University of Illinois at Chicago, Minqing Hu and Bing Liu. It is composed of two lists of words: 2006 positive keywords and 4783 negative ones.

Since the task targets hate speech, we have used only the list of negative words. Experimenting on the training set, we have identified the minimal optimal number of keywords to appear in a tweet, so that it is considered to contain hate speech. Our experiment showed that this minimal number is 4: Every tweet with four or more negative words were labeled as containing hate speech.

After manually inspecting the training set, we have also identified few entities which were strongly associated with hate speech inside the training and evaluation corpora (one of them "Greta Thunberg") and added them to the lexicon.

5 Results and discussion

We have participated in Sub task A, whose goal was to detect from the test set the tweets, containing hate speech. Our system ranked 18 out of 22 participating systems, with F1 score of 0.83. (F1 was the official ranking criteria of this shared task). Our score was 0.03 lower than the system in the middle of the ranking. We have obtained a score only 0.01 lower than the preceding in the ranking system, which used a large language model; moreover, our precision was among the highest ones.

Considering our accuracy, we ranked 13, which is caused by the high precision of the rule based approach and the prevalence of instances, belonging to the negative category (no hate speech). Our accuracy was also higher than the accuracy of the established baselines for this task, reported in (Shiwakoti et al., 2024).

Table 1 displays examples of hate speech tweets identified by our system. Notably, the detection of a substantial number of hate speech tweets was facilitated by the presence of the named entity "Greta Thunberg", which we had identified as a hate speech indicator in the training set. However, it's important to note that this observation reflects a specificity of the shared task data rather than a broader trend on Twitter.

Moreover, refining the focus on tweets containing a high number of negative keywords proved to be an effective strategy for achieving high precision in hate speech detection.

6 Conclusions

We introduced a lexicon-based system designed to identify hate speech in tweets related to climate change. Despite its simplicity and orientation towards high precision, our system achieved accuracy above the baseline and F1 score comparable to some machine learning approaches. Our lexicon-based method achieved one of the highest precision scores of 0.92.

However, it ranked in the low part of the leaderboard, primarily attributed to its notably low recall of 0.777. This was due to the simplicity of our approach with respect to other lexicon based works. We invested relatively little time in its development, which did not allow us to exploit the full potential of this class of methods.

References

MA Al-Shabi. 2020. Evaluating the performance of the most important lexicons used to sentiment analysis and opinions mining. *IJCSNS*, 20(1):1.

Venkateswarlu Bonta, Nandhini Kumaresh, and N Janardhan. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Christopher SG Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

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.

Outi Niininen and Stefan Baumeister. 2022. 12 fridays for future wants to save the world—but what do people think about the movement? *Social Media for Progressive Public Relations*, page 66.

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.

Michel Rosenfeld. 2002. Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo L. Rev.*, 24:1523.

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*.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, Marco La Cascia, et al. 2021. Detection of hate speech spreaders using convolutional neural networks. In *CLEF (Working Notes)*, pages 2126–2136.

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)*.