Expose: Transformer Methods and their Biases on German Hate Speech Detection Datasets

Isadora White

August 2022

1 Motivation

Hate speech is a phenomenon on online platforms such as Facebook and Twitter as well as in the online comment sections of newspapers that can disrupt and derail otherwise productive and respectful online debates [1]. Facebook, Twitter and newspapers such as die Rheinische Post and the New York Times [15] have all employed automation techniques to combat hate speech on their platforms. However, the automated moderation problem is far from trivial. Automated algorithms make mistakes, can be biased in harmful ways and also be mistrusted by users [10]. The problem is made even more difficult for platforms where the majority of users comment in languages other than English, since language models developed to perform well on English often do not perform as well on other languages [1]. In this paper we will analyze an application of T5, a state of the art transformer model, to the problem of German hate speech detection and investigate the biases of this model and the datasets it was trained on.

On the German hate speech detection datasets RP-Mod [1] and the dataset GermEval-2018 [17] the machine learning model which achieves the highest f1 score is GBERT [2] a version of the model BERT [6] trained on the German language. Similarly, the multi-lingual version of BERT [6] named XLM RoBERTa [3] has achieved the high performance on hate speech detection datasets in Hindi [12] and Arabic.

T5 [11] is a transformer model based on the original encoder-decoder model for transformers [16]. In an English language setting, T5 has achieved state of the art performance [13] on the Hate Speech Detection datasets HASOC 2021 [9] and OLID (OffensEval 2019, Task 6 of SemEval 2019) which both consist of tasks related to a binary classification of hate speech. We hypothesize that T5 trained on a German corpus could also achieve state of the art performance on German language hate speech detection datasets.

However, state-of-the-art performance on hate speech detection datasets does not imply flawless hate speech detection. There is evidence that hate speech classifiers are over-attentive to group identifiers such as "mexican", "homosexuality", and "islam" when trained on English hate speech datasets [4]. Such biases could lead to harmless comments containing these words being unnecessarily censored and performance on real-world datasets to decreasing.

The most likely cause of such biases is due to unbalanced datasets in which the group identifiers appear more frequently in problematic (positive) comments than in unproblematic comments. In a study done of several English language datasets there was evidence of racial bias detected in all of the datasets tested [5]. Classifiers trained on these datasets predicted that tweets written in a style which indicated an African-American background were predicted to be problematic at substantially higher rates. [5]

We hypothesize that German language datasets have similar, if not identical, biases to the English datasets, and that group identifiers also play a large role in the incorrect classification of the models.

2 Formulation of Goals

The first goal of the proposed paper is to add a German T5 and its variations to the arsenal of transformers which can be easily applied to the German language hate speech detection datasets and thereby provide improvements over the current SoTA models in terms of binary f1 score and accuracy. We will also test our fine-tuned German T5 on a variety of German datasets to validate the results and demonstrate that T5 can be applied generally. After the publication of this paper and the associated code repository, it should be easy for users to deploy our fine-tuned German T5 model for hate speech detection.

For us, it is not enough to have an easily deployable state of the art transformer model. It is also imperative that we analyze the potential biases of our model and the datasets they were trained on. We would like to check if there are group-identifiers in the German language datasets which are over-attended by the models and datasets as they are in English language datasets [4]. In addition to over-attending to specific group identifiers, it is possible that the model will flag comments written in a different style or slang of German to be problematic. For example, linguistic phenomena such as grammatical errors or a more informal style may be flagged more frequently as problematic. Thus, the second goal of this paper is to uncover these biases in the various different German language datasets we test.

3 Research Methodology

The following datasets will be used in our experiments:

- GermEval-2018 [17] is a hate speech detection dataset labeled by one of the three organizers of the task and there was a deliberate attempt by the authors to debias the dataset by sampling further words containing words such as ”Merkel” or ”Fluchtlinge”.

- RP-Mod, RP-Crowd-2, and RP-Crowd-3 [1] are sourced from comments made on newspaper articles published by the Rheinische Post. RP-Mod is a dataset where the comments are labelled by moderators and RP-Crowd-2, RP-Crowd-3 are datasets labelled by crowdworkers.

- Der Standard [14] consists of user comments posted to the website of a German-language Austrian newspaper. Professional forum moderators have annotated 11,773 posts according to seven categories they considered crucial for the efficient moderation of news articles. There are also a million unlabeled examples in the Der Standard dataset.

- HASOC [9] is a dataset of tweets in Hindi, English, German, and Marathi classified into "HOF" (hate or offensive) or "NOT" (not hate or offensive). For this paper we will utilize the German dataset for the evaluation of our models.

For each dataset, we will compare f1 and accuracy scores of GBERT [2], T5, and XLM RoBERTa [3] and then come to a conclusion as to which model performs the best on German hate speech detection datasets. By evaluating the models on four diverse datasets, we will do an effective comparison of the transformer models on hate speech detection datasets and provide more conclusive results.

Then, using the novel German T5 model we will analyze the biases of the dataset using the following methods:

- SHAP values [8]: apply SHAP to false positive predictions to assess which words contributed most to the incorrect prediction on a given comment

- SOC values [7]: a hierarchical sequence attribution method which ranks which words contribute the most to positive and negative predictions of a given comment

- Assessment of models on datasets containing counterexamples to common misconception of the models. Where "common misconception" could be for example, an assumption by the model that all comments with the word "Migranten" are positive/problematic and a counterexample would be a comment which is not problematic but contains the word "Migranten" as well

After completing these experiments we should have an assessment of if and how the datasets are biased. This will be represented in the form of a list of words or phrases which the model frequently classifies incorrectly. It is likely that the biases will vary across the datasets due to the different sampling techniques.

By analyzing the biases in the datasets and models we will quantify the existing biases in the datasets and have a comparison to the biases present in English language datasets. Knowledge of the existing biases will help develop solutions and ways of mitigating these biases when these models are used to moderate newspaper or social media comments.

4 Bibliography

References

[1] Dennis Assenmacher et al. *RP-Mod & RP-Crowd: Moderator-and Crowd-Annotated German News Comment Datasets.* Tech. rep. URL: `https://rp-online.de`.

[2] Branden Chan, Stefan Schweter, and Timo Möller. "German's Next Language Model". In: (Oct. 2020). URL: `http://arxiv.org/abs/2010.10906`.

[3] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: (Nov. 2019). URL: http://arxiv.org/abs/1911.02116.

[4] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. *Racial Bias in Hate Speech and Abusive Language Detection Datasets*. Tech. rep. 2019, pp. 25–35. URL: https://www.perspectiveapi.com.

[5] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets". In: (May 2019). URL: http://arxiv.org/abs/1905.12516.

[6] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. 2019.

[7] Xisen Jin et al. *TOWARDS HIERARCHICAL IMPORTANCE ATTRIBU-TION: EXPLAINING COMPOSITIONAL SEMANTICS FOR NEURAL SEQUENCE MODELS*. Tech. rep. URL: https://inklab.usc.edu/hiexpl/.

[8] Scott M Lundberg, Paul G Allen, and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. Tech. rep. URL: https://github.com/slundberg/shap.

[9] Sandip Modha et al. "Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech". In: *ACM International Conference Proceeding Series*. 2021. DOI: 10.1145/3503162.3503176.

[10] Kilian Müller et al. "Exploring Audience's Attitudes Towards Machine Learning-based Automation in Comment Moderation". In: *Proceedings of the 17. Internationale Tagung Wirtschaftsinformatik (WI 2022)*. Publication status: In press. Nürnberg, Germany, 2022. URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=wi2022.

[11] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: (Oct. 2019). URL: http://arxiv.org/abs/1910.10683.

[12] Sayar Ghosh Roy et al. "Leveraging Multilingual Transformers for Hate Speech Detection". In: (Jan. 2021). URL: http://arxiv.org/abs/2101.03207.

[13] Sana Sabah Sabry et al. "HaT5: Hate Language Identification using Text-to-Text Transfer Transformer". In: (Feb. 2022). URL: http://arxiv.org/abs/2202.05690.

[14] Dietmar Schabus, Marcin Skowron, and Martin Trapp. "One million posts: A data set of German online discussions". In: *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, Aug. 2017, pp. 1241–1244. ISBN: 9781450350228. DOI: 10.1145/3077136.3080711.

[15] Shan Wang. *The New York Times, with a little help from automation, is aiming to open up most articles to comments*. Accessed September 17, 2017. https://www.niemanlab.org/2017/06/the-new-york-times-with-a-little-help-from-automation-is-aiming-to-open-up-most-articles-to-comments/. 2017.

[16] Ashish Vaswani et al. “Attention Is All You Need”. In: (June 2017). URL: `http://arxiv.org/abs/1706.03762`.

[17] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. “Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language”. In: *GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)* September (2018).