Dialectic Bias in Toxicity Detection of Google's Perspective API

A study with five parallel corpora

Hanxin Xia | 3417418
hanxin.xia@uni-duesseldorf.de

Abstract

Hate Speech detection has been a heated topic in NLP/NLU since the spread of social media. Despite the existence of various automatic toxicity scoring models, their biases against language varieties that align with minority identities have been discussed by researchers (Sap et al., 2019). In this work, we focus on Google's Perspective API and analyze its toxicity scores on four synthetic corpora with different dialect features compared to an original master corpus. We discovered that Perspective scored all parallel corpora significantly differently than the master corpus, which further confirms the existence of the biases. Moreover, we discovered the capping phenomenon with the toxicity scores of the gold toxic sentences: Gold non-toxic sentences are more liketly to suffer from toxicity score increase, when converted to a dialect, compared to gold toxic sentences.

1 Introduction

With the increasing popularity of social media and the daily widening user bases, the moderation of the contents on online communication platforms have been a crucial part of maintaining the friendliness for the users and building healthy communities. In pursuit of this objective, the platforms have implemented various means including reactive methods such as community guidelines, user report and post-removal of harmful contents and proactive methods either with sensitive words and language pattern rules (Gitari et al., 2015) or machine learning algorithms for automated hate speech detection (Alrehili, 2019).

However current detoxifying approaches are proven to be not perfect. While the reactive methods are commonly subject to deficiency and passivity, the automatic proactive methods often discriminate against minority aligned language varieties other than Standard American English (SAE) (Sap et al., 2019; Zhou et al., 2021).

This research aims to test the performance of Google's Perspective API on five parallel corpora. Despite being a continuously improving model on toxicity scoring (Google, 2024), the bias against non-standard English is still present in the results. The model scored (state: March 2024) texts from the four English dialects: African American Vernacular English (AAVE), Nigerian and Indian dialects and Singaporean English (Singlish). significantly differently than the original texts of standard English. Thus, we posit that there is a continued need for optimization in the hate speech detection of online contents for better mitigating biases that may infringe upon the online representation of individuals from marginalized communities.

2 Background

The process of detoxification and hate speech detection is aimed to create healthy and welcoming online communities. Yet the biases in these processes are leading to not only unfair penalties for minority individuals (Davidson et al., 2019) but also potential complete exclusion of their voices in online spaces (Blodgett and O'Connor, 2017) simply because of the usage of non-standard language.

The underlying reasons behind such biases can be traced back to multiple factors. As Xia et al. (2020) points out, first-hand annotators are inclined to classify non-abusive AAE texts as toxic, which may lead to bias already at the corpus creation stage. Secondly, surface markers on the language such as mention of specific identity terms and usage of certain uncommon patterns are also prone to cause the so-called "key word bias" in models (Resende et al., 2024; Schäfer, 2023). The lack of data of English language varieties in the training data may lead to their incapability to capture the real intentions and deep meanings of such speeches. Last but not the least, suboptimal performances of hate speech detection models can also be attributed to the ignorance of linguistic subtleties engrained in contextual factors such as speaker identity, environment and even special spelling (Davidson et al., 2019).

Several methodologies have been proposed to mitigate these biases. Schäfer (2023) experimented with transformer based multi-class prediction in order to establish a more comprehensive viewpoint for hate speech detection with less success. Moreover, the approach of masking out identity terms exhibits obvious drawbacks despites its merits. By raising subjectivity level when identity terms occur, (Zhao et al., 2022) were able to efficiently neutralize their model's bias against minority aligned texts. Similar effects can be achieved by including confident dialect prediction (Ball-Burack et al., 2021). By implementing adversarial learning during the model learning process, Xia et al. (2020) and (Okpala et al., 2022) were both able to reduce baseline model's false positive rate facing African American English like texts.

3 Data

This section provides an overview on the dataset used in this research. The data augmentation and synthesizing technique will be explained. The choice of the synthetic approach will also be substantiated.

3.1 HateXplain

The data the research is based on stems from the publicly available HateXplain[1] dataset. HateXplain is designed for interpretation of hate speech detection results based on certain text features. It includes 20,047 instances from Twitter and Gab posts collected around the year 2020. Besides the class labeling (*hate*, *offensive* or *normal*), targeted groups and rationales (text chunk that directly contributes to the decision) are also included in the annotation scheme. Each post is annotated by three Amazon Mechanical Turk (MTurk) workers. 919 instances, where all three annotators decide for a different class, are excluded from the final compilation, which ensures the credibility of the annotation results (Mathew et al., 2021).

Since the original class label is tri-partie and the dichotomy of toxicity is interested in the research, we combine *hate* and *offensive* together as toxic, and reinterpret *normal* as non-toxic. Based on majority voting among the three annotators, we are able to retrieve 12,276 toxic and 7,771 non-toxic posts.

The HateXplain dataset with the newly interpreted labels will be referred to as the **original** or **master dataset** in the following texts.

3.2 Synthetic datasets

Acknowledging the restricted capabilities of current NLP systems on different English language varieties, the Multi-VALUE[2] package (Ziems et al., 2023) provides a complete framework for rule-based English dialect transformation and evaluation. By applying syntactic and morphological transformation rules based on 189 unique features, Multi-VALUE is able to convert SAE to 50 English dialects.

In this experiment, we choose four representative dialects: African American Vernacular English (**AAVE**), Nigerian English (**NigerianD**), Indian English (**IndianD**) and Colloquial Singapore English (**Singlish**). The selection includes English dominant communities spanning across three major continents with different historic relations with the English language, which provides good representativeness on the general tendency comparing standard English and dialects.

[1] `https://github.com/hate-alert/HateXplain`

[2] `https://github.com/SALT-NLP/multi-value/tree/main`

3.3 Data choice

4 Methodology

5 Results

6 Conclusion

References

Ahlam Alrehili. 2019. Automatic hate speech detection on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6.

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

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.

Google. 2024. Perspective api. Accessed: 2024-06-19.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. Aaebert: Debiasing bert-based hate speech detection models via adversarial learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1606–1612. IEEE.

Guilherme H Resende, Luiz F Nery, Fabrício Benevenuto, Savvas Zannettou, and Flavio Figueiredo. 2024. A comprehensive view of the biases of toxicity and sentiment analysis methods towards utterances with african american english expressions. *arXiv preprint arXiv:2401.12720*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Johannes Schäfer. 2023. Bias mitigation for capturing potentially illegal hate speech. *Datenbank-Spektrum*, 23(1):41–51.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2022. Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Social Networks and Media*, 29:100205.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Code Access

Source code and detailed instructions for conducting analyses are publicly accessible under: https://github.com/xuanxuanx-98/ToxBias-ParallelCorpora.