

Northumbria Research Link

Citation: Sinclair, Mark and Gabriel, Aaron (2022) Explainable Supervised Models for Bias Mitigation in Hate Speech Detection: African American English. In: TEI 2022 Conference Book: 12-16 September 2022, Newcastle University, Newcastle upon Tyne. Newcastle University, Newcastle upon Tyne, TEI2022 -91.

Published by: Newcastle University

URL: <https://doi.org/10.5281/zenodo.7120027>
<<https://doi.org/10.5281/zenodo.7120027>>

This version was downloaded from Northumbria Research Link:
<https://nrl.northumbria.ac.uk/id/eprint/50917/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

**Explainable Supervised Models for Bias Mitigation in Hate Speech Detection:
African American English.**

Aaron Gabriel

Northumbria University

Dr Mark Sinclair

Northumbria University

Abstract

Automated hate speech detection systems have great potential in the realm of social media but have seen their success limited in practice due to their unreliability and inexplicability. Two major obstacles they have yet to overcome is their tendency to underperform when faced with non-standard forms of English and a general lack of transparency in their decision-making process. These issues result in users of low-resource languages (those that have limited data available for training) such as African-American English being flagged for hate speech at a higher rate than users of mainstream English. The cause of the performance disparity in these systems has been traced to multiple issues including social biases held by the human annotators employed to label training data, training data class imbalances caused by insufficient instances of low-resource language text and a lack of sensitivity of machine learning (ML) models to contextual nuances between dialects. All these issues are further compounded by the 'black-box' nature of the complex deep learning models used in these systems. This research proposes to consolidate seemingly unrelated recently developed methods in machine learning to resolve the issue of bias and lack of transparency in automated hate speech detection. The research will utilize synthetic text generation to produce a theoretically unlimited amount of low-resource language text training data, machine translation to overcome annotation conflicts caused by contextual nuances between dialects and explainable ML (including integrated gradients and instance-level explanation by simplification). We will attempt to show that when repurposed and integrated into a single system these methods can both significantly reduce bias in hate speech detection tasks whilst also providing interpretable explanations of the system's decision-making process.