Hate-speech detection on social media

Machine Learning and Artificial Intelligence
Berkely Eng, Berkely Hass. Nov 2022.

Hagay Zamir

June 7, 2023

Abstract

This project focuses on the challenges faced in automatic hate-speech detection on social media, particularly the difficulty in distinguishing hate speech from other forms of offensive language. Traditional lexical detection methods tend to have low precision, as they classify any messages containing specific terms as hate speech. Previous supervised learning approaches also struggled to differentiate between the two categories effectively. To address this, a crowd-sourced hate speech lexicon was utilized to collect tweets containing hate speech keywords. A multi-class classifier was trained using this dataset, and close analysis of the predictions revealed instances where hate speech could be reliably separated from other offensive language, as well as situations where differentiation proved more challenging. The findings indicated that racist and homophobic tweets were more likely to be classified as hate speech, while sexist tweets were predominantly labeled as offensive. Tweets lacking explicit hate keywords posed additional difficulties in classification. This project provides valuable insights into the complexities of hate-speech detection and underscores the significance of thorough analysis for enhancing classification accuracy.

Introduction

Defining hate speech and differentiating it from offensive language is a complex task without a formal consensus. Generally, hate speech refers to speech that targets disadvantaged social groups in a potentially harmful manner. In the United States, hate speech is protected under the First Amendment but has been extensively debated in legal and campus speech code contexts. Other countries like the United Kingdom, Canada, and France have laws prohibiting hate speech, often defined as speech that targets minority groups and could incite violence or social disorder, carrying penalties such as fines and imprisonment. Online platforms like Facebook and Twitter have implemented policies to curb hate speech by prohibiting attacks based on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence. Our definition of hate speech encompasses language expressing hatred, derogation, humiliation, or insults towards targeted groups, including extreme cases that involve threats or incitement to violence. However, our definition does not encompass all instances of offensive language, as certain terms may be used differently within specific communities without the same intent of hate. To address the conflation of hate speech and offensive language, we labeled tweets into three categories: hate speech, offensive language, or neither. Using this dataset, we trained a model to distinguish between these categories and conducted a detailed analysis to understand the challenges involved in accurate classification. The results highlight the importance of fine-grained labeling and the need to consider context and the diverse usage of hate speech in future research.