Understanding Implicit Hate Speech Detection

Master/Bachelor Thesis

Motivation

Implicit hate speech is defined by coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity. Compared to explicit hate speech detection, implicit hate speech contains several challenges for the NLP models. One key challenge, is that implicit hate speech detection does not contain clear lexical flags like profanity or swear-words. In addition, sometimes it might also contain a "positive" sentiment due to linguistic phenomena like sarcasm, humour, euphism etc. In this thesis we would want to address the implicit versus explicit hate speech detection, by comparing the lexical cues used in the text, and also extending with information from a) the author of the hate text, and b) the targe of the hate text.

Difficulty

Analysis

Programming

Literature

Contact

Prof. Dr. Lucie Flek
Joan Plepi

joan.plepi@uni-marburg.de
caisa-lab.github.io
Hans-Meerwein-Straße 6, 35043, Marburg
Room: 04C21 (Staircase C, Level 4)

Task Description

This thesis will start with analysis of one of the datasets (or both) Latent Hatred benchmark corpus from Twitter [1] or TOXIGEN [2], that is a machine generated dataset. After analysing and understanding more about implicit hate speech taxonomy compared to explicit one, we would like to build some machine learning models to compare between both. In addition, we will extend the Latent Hatred dataset from [1], with additional tweets, and their author and audience information. We would like to analize author social network, and model the targeted audience to whom the hated speech is directed, in order to analyze the effect of these features in implicit hate speech detection.

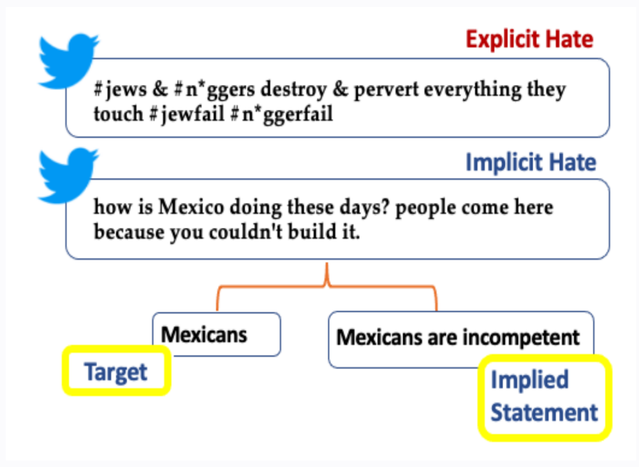

Figure 1: A sample post of explicit vs implicit hate speech taken from [1]

References

[1] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.