HateSet: A dataset for Hate Speech Detection

Akshay Atam
aatam@stevens.edu

Vikrant Gajare
vgajare1@stevens.edu

Introduction

Hate crimes are unluckily not anything new in society. However, social media and different means of on-line verbal exchange have all started playing a bigger role in hate crimes. For instance, suspects in numerous current hate-related terror assaults had an intensive social media history of hate related posts, suggesting that social media contributes to their radicalization [1].

Hate speech is an act of communication that is intended to harm or intimidate a person or group based on their membership in a certain social category. It is often characterized by a message of intolerance or bigotry. Hate speech can take many forms, including but not limited to, name-calling, slurs, offensive jokes, and hate symbols. Social media has enabled users to express their views freely, at times, with anonymity. While the ability of free speech is a human right that should be cherished, spreading hate and violence towards a certain group is an abuse to this liberty.

Background and Related Work

Collecting and annotating hate speech data for training classifiers is hard and cumbersome. Specifically, identifying and agreeing whether or not a specific text should be labelled as hate speech is tough, as formerly mentioned, there is no generic definition of hate speech. One example of hate speech data set is “Dynahate” [4]. This dataset consists of hate speech text which is generated and labelled by trained annotators over four rounds of dynamic data creation. Another good example is data collected from twitter. This data can be used to detect hate speech in tweets, to check whether a tweet has a racist or sexist sentiment associated with it.

Dataset Description

Our work proposes a formulation of a novel dataset, outsourced from existing datasets and leveraging the power of GPT-3 to construct the test cases. However, one cannot just generate test cases based on what he/she thinks is described as hate speech. We will be using the Facebook Community guidelines [5] as a base to construct hate speech data. There are three tiers in the Facebook community standards, each describing in detail of what constitutes as hate speech.

There are a lot of datasets available for hate speech detection but we cannot find any single dataset that encapsulates a majority of what is considered to be hate speech. It is a very subjective topic. The use of Facebook community guidelines has been chosen because it encapsulates a lot of topics that relate to hate speech.

Research Question

There is a relationship between a strong model and a strong dataset. The aim of the dataset is to have such test cases that the model fails to detect its true label. We can create a strong model that has an impressive accuracy on one dataset. However, when tested against unseen data of the same dataset or a completely new dataset, it fails to provide that accuracy.

Thus, our research question is how can we create a dataset that has such training examples that a machine learning model, if trained, provides great accuracy not only in training but in testing as well.

Plan for experiment

Vikrant and I will be generating the dataset according to, as formerly mentioned, the community standards set by Facebook. We would be generating 50 test cases for each of the policies mentioned in the Facebook community standards policy. There are three tiers of policies and we would be working in the following manner:

1. Tier 2 requires generating test cases that are generic and could be created manually and taken from existing datasets. We will begin our test case generation from there.
2. Tier 1 requires generation of test cases that are specific to certain ethnic groups. Since publicly available datasets might not have hate speech targeted to certain groups, we would have to either generate test cases manually or use GPT-3 this task.
3. Tier 3 involves generation of test cases on the basis of people's protected characteristics. This tier is the most subjective of the three and would require creation of specific test cases.

We have prioritized speed of creating test cases over the complexity of the cases to ensure that we get the most work done in the initial stage of our progress. Completion of tiers 1 and 2 would allow us to be familiar with test case generation and thus facilitate creativity.

References

[1] MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, et al. (2019) Hate speech detection: Challenges and solutions. PLOS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152

[2] Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: The 3rd Workshop on Natural Language Processing for Computer-Mediated Communication @ Conference on Natural Language Processing; 2016.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[4] Singh, A. (n.d.). APS/dynahate · datasets at hugging face. aps/dynahate · Datasets at Hugging Face. Retrieved November 25, 2022, from https://huggingface.co/datasets/aps/dynahate

[5] https://transparency.fb.com/policies/community-standards/hate-speech/