Katz

Katz School
of Science and Health

Yeshiva University - Katz School of Science

Data Analytics and Visualization MS

date: February 2022
from: Jesus Olivera, Independent Study

subject: YU Hate Speech Project

Designing and Implementing a Data Labeling Tool for ML/AI Hate Speech Classification

Introduction

In recent years, American Jews have faced increased threats of violence and harassment off and online. According to ADL (Anti-Hate Crime Organization) Antisemitic incidents in 2019 and 2020 were, respectively, the highest and third-highest years on record for cases of harassment, vandalism, and assault against Jews in the United States since the organization started tracking these incidents back in 1979.

In the past few years , tech platforms have focused resources on trying to combat online hate speech. This project is designed to help advance the efforts around detecting and preventing Antisemitic online hate speech by providing Yeshiva University’s Research Department a framework for data pre-processing and data flows for the development of Machine Learning and AI capabilities aimed at detecting and preventing online Antisemitic hate speech.

Problem Statement

Antisemitism is on the rise. New technologies and applications are required to combat the negative effects of online hate speech towards the Jewish Community. For this reason, a complex generic use case for an MVP aimed at designing, building and deploying an Antisemitic data labeling pipeline to feed Machine Learning and AI algorithms that detect hate speech towards Jews will be done.

Objective

The objective of this study is to organize, guide, and design a development and implementation strategy for an MVP around detecting online hate speech and building a robust antisemitic labeled dataset to train ML and AI algorithms.

Katz
Katz School
of Science and Health

Importance & Broader Impact

Developing techniques and methods for combating online hate speech can provide tech companies with a framework around monitoring content in their platforms. Overtime, these initiatives can enhance regulatory compliance efforts and ultimately provide users a safe space for their regular interactions within online platforms.

Mental health has declined over the years and some of the variables that have impacted the overall uptake in suicides, depression and other mental health problems can be attributed to the rise in technology and in particular social media.

A future state for this project can include providing access to external organizations, via a data marketplace or a deployment space, to high-quality antisemitic labeled datasets for their models and/or access to trained models focused on detecting antisemitism online.

This project has the potential of contributing to the overall benefit of society by mitigating some of the negative effects of online hate speech towards Jews and the broader communities.

Katz
Katz School
of Science and Health

YU Hate Speech Backend Engineering Focused MVP Design

Sprint 1

High-Level Project Development Phase 1:

Area of Focus: Backend

Timeframe: January 2022 to May 2022

Owner: Jesus Olivera, Lead Engineer

- Literature and Existing Components Review
- Gap Analysis
- Tools Selection
- Process Logic Development
- Architecture Design
- Pipelines Design & Deployment Instructions

Sprint 2

High-Level Project Development Phase 2:

Area of Focus: ML / AI

Timeframe: May 2022 to TBD

Owner: TDB

- Model Design
- Model Development
- Model Deployment
- Model Inputs & Outputs Process Logic Design, Development & Implementation
- Model Setup

Sprint 3

High-Level Project Development Phase 3:

Area of Focus: User Interface

Timeframe: TBD

Owner: TDB

- Design Thinking: Understanding users and their values
- Competition Research
- Interface Sketch/ Wireframe
- Interface Design
- Design Implementation
- Implementation Evaluation

Katz

Katz School
of Science and Health