Hate speech detection using AI

-Project by Harshith, cs20b1123

Introduction and motivation to take up the project :

Hate crimes are unfortunately nothing new in society. However, social media and other means of online communication have begun playing a larger role in hate crimes. For instance, suspects in several recent hate-related terror attacks had an extensive social media history of hate-related posts, suggesting that social media contributes to their radicalization

Vast online communication forums, including social media, enable users to express themselves freely, at times, anonymously. While the ability to freely express oneself is a human right that should be cherished, inducing and spreading hate towards another group is an abuse of this liberty. As such, many online forums such as Facebook, YouTube, and Twitter consider hate speech harmful, and have policies to remove hate speech content .

Due to the societal concern and how widespread hate speech is becoming on the Internet , there is strong motivation to study automatic detection of hate speech. By automating its detection, the spread of hateful content can be reduced.

Challenges in solving this problem ?

1. Automatic hate speech detection is technically difficult;
2. Some approaches achieve reasonable performance;
3. Specific challenges remain among all solutions;
4. Without societal context, systems cannot generalize sufficiently.

Defining hate speech

The definition of hate speech is neither universally accepted nor are individual facets of the definition fully agreed upon. A clear definition of hate speech can help the study of detecting hate speech by making annotating hate speech an easier task, and thus, making the annotations more reliable. However, *the line between hate speech and appropriate free expression is blurr*..

1. Encyclopedia definition : "Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity."
2. Twitter: "Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."
3. Facebook: "We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation."

It is notable that in some of the definitions above, a necessary condition is that it is directed to a group. This differs from the Encyclopedia , where an attack on an individual can be considered hate speech. ***A common theme among the definitions is that the attack is based on some aspect of the group or peoples identity***.

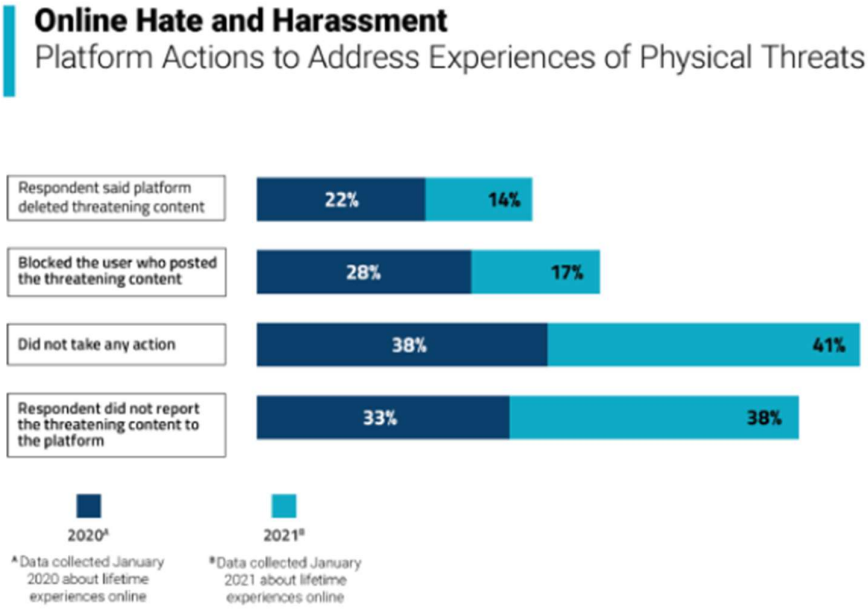
1. Hate speech is to incite violence or hate
2. Hate speech is to attack or diminish
3. Hate speech has specific targets
4. Whether humor can be considered hate speech

Let's look at an example to understand the thin line between hate speech and a normal statement:

A particular problem not covered by many definitions relate to factual statements. For example, "Jews are swine" is clearly hate speech by most definitions (it is a statement of inferiority), but "Many Jews are lawyers" is not. In the latter case, to determine whether each statement is hate spee ch, we would need to

check whether the statement is factual or not using external sources. This type of hate speech is difficult because it relates to real-world fact verification—another difficult task. More so, to evaluate validity, we would initially need to define precise word interpretations, namely, is "many" an absolute number or by relative percentage of the population, further complicating the verification.

Datasets



Social media platforms are a hotbed for hate speech, yet many have very strict data usage and distribution policies. This results in a relatively small number of datasets available to the public to study, with most coming from Twitter (which has a more lenient data usage policy). While the Twitter resources are valuable, their general applicability is limited due to the unique genre of Twitter posts; the character limitation results in terse, short-form text. In contrast, posts from other platforms are typically longer and can be part of a larger discussion on a specific topic. This provides additional context that can affect the meaning of the text.

We will be using Kaggle provided dataset and will also try to take msgs from twitter live for this project.

- **Kaggle** : Kaggle.com hosted a shared task on detecting insulting comments. The dataset consists of 8,832 social media comments labeled as insulting or not insulting. While not necessarily hate speech, insulting text may indicate hate speech.

Approach and use of AI :

Most social media platforms have established user rules that prohibit hate speech; enforcing these rules, however, requires copious manual labor to review every report. Some platforms, such as Facebook, recently increased the number of content moderators. Automatic tools and approaches could accelerate the reviewing process or allocate the human resource to the posts that require close human examination.

Machine learning models/classifiers take samples of labeled text to produce a classifier that is able to detect the hate speech based on labels annotated by content reviewers. Various models were proposed and proved successful in the past. We describe a selection of open-sourced systems presented in the recent research.

Hate speech detection approach I look to use :

Naïve Bayes, Support Vector Machine and Logistic Regression.

These models are commonly used in text categorization. Naïve Bayes models label probabilities directly with the assumption that the features do not interact with one another. Support Vector Machines (SVM) and Logistic Regression are linear classifiers that predict classes based on a combination of scores for each feature. Open-source implementations of the these models exist, for instance in the well-known Python machine learning package *sci-kit learn* .

Naïve Bayes models work on the assumption that the features are independent of each other, which means that the probability of a particular feature being present does not depend on the presence or absence of other features. This makes Naïve Bayes models computationally efficient and easy to implement.

Support Vector Machines (SVM) and Logistic Regression are linear classifiers that use a combination of scores for each feature to predict the class of a given text. SVMs work by finding a hyperplane that separates the data into two classes, while Logistic Regression uses a logistic function to predict the probability of the input text belonging to a particular class.

To use these models for hate speech recognition, you would first need to train the models using a labeled dataset of hate speech and non-hate speech texts. The models would then use this training data to learn how to distinguish between hate speech and non-hate speech texts based on various features such as the presence of certain keywords, the use of profanity, or the use of derogatory terms.