SEMESTER 4

Fontys University of Applied Science

AI PROJECT PROPOSAL

KARINA KOZAROVA

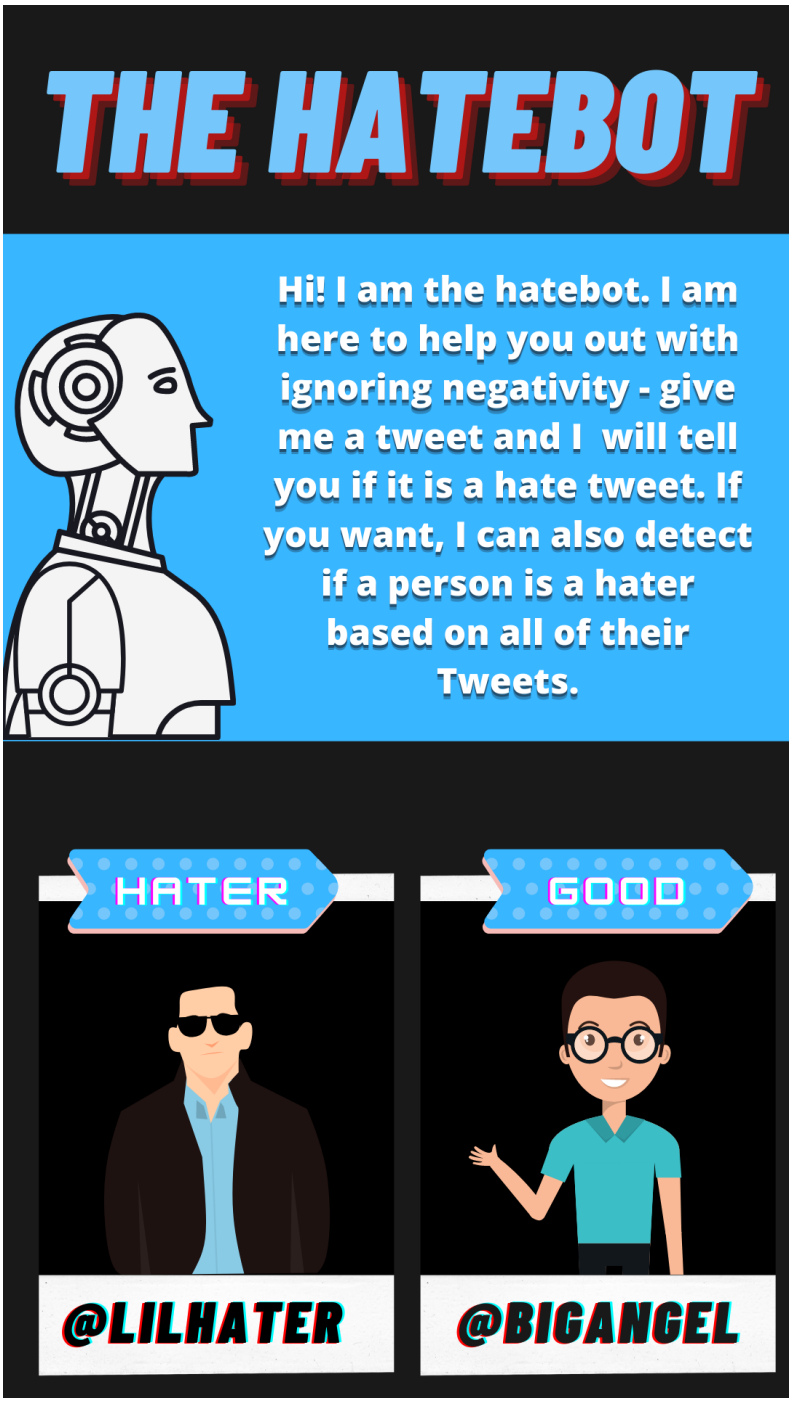
ARTIFICIAL INTELLIGENCE SPECIALIZATION. ENGLISH STREAM.

Table of Contents

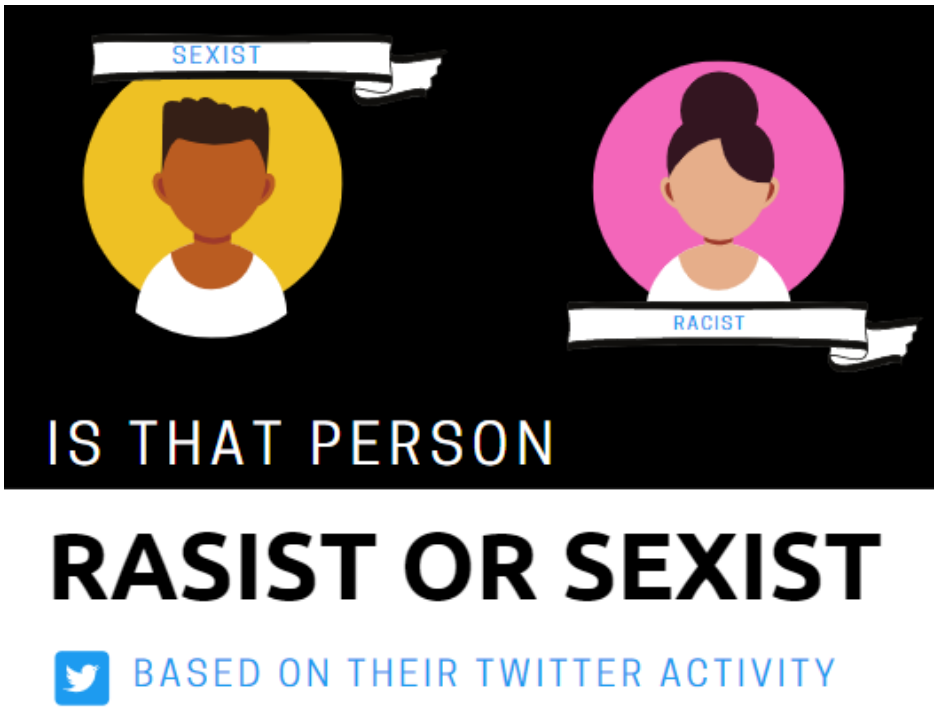
Table of Contents	2
The idea	3
Social relevance and the need for a hate bot	4
Target users	6
Technical approach	7
Datasets to use	7
Summary	7

The idea

Negativity is all around us - from the news on the TV through our social networks and even our conversations with family and friends. What if we want to avoid that at least on social networks where some posts are simply hate speech, they don't add any kind of value to our life? Well, with the hatebot we can detect if a Tweet is using hate speech or offensive language. Once we know that, the user can decide for themself if they want to read a tweet (this way not affecting anyone's right to express their opinion no matter what it is while still giving the user the choice to filter out negativity). Also, the hatebot gives us the option to analyze all the tweets a person has created and classify the person as a hater or not thus allowing the user to choose if they want to affiliate with that user.



Social relevance and the need for a hate bot



Bullying. One of the biggest issues that all youth face. 59% of U.S. teens have been bullied or harassed online. The vast majority of teens believe online harassment is a problem that affects people their age, and 63% say this is a major problem. 7 in 10 young people experience cyberbullying before they hit the age of 18.[1] Those statistics are frightening to say the least.

However, the problem is bigger than that, bullying is not something only the young people have to deal with - even people like the ex-President of the United States of America, a position that is probably one of the most important in the world, Donald Trump is doing it and dealing with it. One of his latest tweets ended with him being blocked from Twitter. In a part of Twitter's official explanation on why they blocked him, we can also see clearly the need for such a bot:

After close review of recent Tweets from the @realDonaldTrump account and the context around them — specifically how they are being received and interpreted on and off Twitter — we have permanently suspended the account due to the ***risk of further incitement of violence****.* [2]

The important part in that statement is the **risk of further incitement of violence**. Social media should be platforms where people can freely and truly express themselves and their opinion. However, sometimes people's opinions do more harm than good. Right now, in such cases usually, the account of the offender either gets suspended or nothing happens. Sometimes we also have the opposite case - perfectly fine tweets can be marked as harmful even though they are not. It is very complicated to understand the difference between intentional hate speech and radical ideas. This also comes from cultural differences. For example, there are still countries where homosexuality is not something accepted so calling someone gay can be hate speech at its finest while in others it can just be used to describe a man who is in love with another man.

[1] According to the following research
https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/

[2] Full statement can be read at https://blog.twitter.com/en_us/topics/company/2020/suspension.html

All social networks are trying to fight hate speech, racism and sexism. Their success of course varies. But the need for something that detects any type of hate speech is there. What makes the hatebot difference than all the other options we currently have is the fact that it does not delete the tweet. Even if it goes over community standards, it can still be viewed but the user is given a caution that it's hate speech.
This way we can make sure that:

- Bullying is slowed down. Users get to ignore people and tweets that might hurt them
- We have less negativity in our lives - less hate speech, more meaningful content
- Accounts that are created for spamming negativity, sexism, racism or bullying are detected and can be stopped
- We understand if something is indeed hate speech or we simply do not like it. However, things aren't always black and white.

However, the hate bot might have a dark side too. Classifying algorithms are racist to begin with - there is not enough data from black people or indigenous for example to be able to train a model as good as if we only looked at white people. This might also influence our hate speech classifier or even make the racist detector racist which can defeat the purpose of the project. This is a complicated issue that araises in all kind of machine learning solutions that we should be aware of. To prevent it, we should make sure that we have data for different types of people and not for example only straight white people living in Los Angeles.

Also, this technology can be used against certain (ethnic) groups and (social) classes. Depending on the data that the bot is trained with, it might become biased thus ruling out a totally normal opinion as hate speech. This is something that the developers should be aware of when developing the solution. The worst possible thing that can happen using the hatebot is having normal opinions marked as hate speech but that is something that is possible with any kind of solution. To prevent it from affecting people, the users should be made aware of that side effect.

To sum up, the need for a system to detect if a tweet is hate speech is there. By developing this solution, we can make sure that there is less hate speech in Twitter users' life thus improving the mental being of the social media users'.

Target users

Job Title
Student

Age
18 to 24 years

Highest Level of Education
Bachelor's degree (e.g. BA, BS

Social Networks

Continent

Europe

Type of social media user

Very active

Wants

Less negativity in her life

More memes, less hate speech

Hey you! My name is Sophie. I am what you might call the classic Twitter user - I post a lot and sometimes get hate speech from people I don't even know. I was born and raised in a European country. Other than Twitter I use a lot of other social media - Facebook, Instagram, even Pinterest. I never got into TikTok and Twitter is still my favourite social media because of the text limits. Daily I spent more than 4 hours on social networks, I want to see more memes and cat videos and less negativity. All the hate speech on Twitter really bothers me, I wish i could ignore it but I always read it. I am currently in a university so most of my time in the lockdown is spent in front of the PC/phone.

Technical approach

To be able to understand if a tweet is hate speech, machine learning will be used. The algorithm will be given a tweet as a text and then will be able to tell us using classification if the tweet is using hate language or not. The classification model will be trained with the datasets mentioned in Datasets. The data we have will be split into test and train data with the ratio of 30 (test) to 70 (train). If the classification model is less than 60% accurate (according to a confusion matrix) and the model can not be tweaked to become better at it, a neural network might be trained.[3]

In the future, the algorithm might be upgraded to also detect if a Twitter user is sexist/racist based on all of his recent tweets. However, at the moment that seems like the labeling is a bit too much unless the algorithm is extremely accurate.

When we are able to classify a tweet as hate speech or not, a website will be built where the user can enter the text of the Tweet and the web application will tell him if it is hate speech or not. In the future another page to the web application will be made where the user can enter a url to a Twitter user's profile and a machine learning algorithm will tell him how many percent the user is using hate speech. However, this requires some more computational power so it might become a premium feature this way allowing the project to be monetized.

Datasets to use

Hate speech dataset: Hate Speech and offensive language dataset
Normal tweets dataset: Twitter tweets dataset
Kaggle dataset - Twitter hate speech
Manually marked by users dataset of offensive tweets - github repo

Summary

Hate speech is something that is a big issue nowadays especially with social media like Twitter. Negativity that comes out of it can take a toll on a person's mental health. To minimize the damage of that, the HateBot can return if a tweet is hate speech or not by being given a Tweet. This way we can ensure that the world becomes if not a better place then at least a place where we can influence the amount of negativity we get.

[3] If I have the time, I would like to work on that even if the model is successful to see if it can be improved that way. However, my estimate is that I won't have the time before the first deadline so