SEMESTER 4

Fontys University of Applied Science

DEPLOYMENT RECOMMENDATION

KARINA KOZAROVA

ARTIFICIAL INTELLIGENCE SPECIALIZATION. ENGLISH STREAM.

Table of Contents

Results of the machine learning

In the machine learning part of the project, a model was created using a SVC with a CountVectorizer and GridSearchCV for hyperparameter tuning. The end results of the model when tested against a part of the dataset is the following:

Accuracy: 0.9515803631472763
Precision: 0.9763303950375449
Recall: 0.9653001936733376

We can see that the results are very high and definitely sufficient for our purpose.

How can the model be applied in the real world?

Now that we have created a successful machine learning model that is able to classify if a Tweet is a hate speech, it is important to explain how that should be deployed.

Possible Implementations

The range of possible implementations about the model is huge, to narrow the field, those are the ones I would pursue:

- An API is present that anyone can (for free) use that returns if a text that is in Twitter's text length limits is a hate speech or not
- A web application where the user can enter the text of the Tweet and the web application will tell him if it is hate speech or not.
- 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.
- Browser plugin that signals the user after they have read a certain number of hate speech tweets per day
- Statistical dashboard that shows how much hate speech

Why not monetize every part of the project?

You probably have already noticed that most parts of this project are available to the public for free and open sourced. From the beginning, the idea of the project was to build something useful that might help (even if only one person) to ignore the negativity on Twitter. I personally do not use Twitter but I have seen the devastating effects the hate speech there can have on people. Because of that, making profit out of this project is not the most important part and thus most of the features will be free to use and totally open source.

Impact analysis

Because of the high level of accuracy, our algorithm is not very prone to errors but those might still occur. Because of this it is important that all of our users are well informed and alerted that even though rarely, sometimes misclassification can happen. However, that is not extremely dangerous because in most cases it won't happen and even when it might happen, the user would still be better off using our application in the first place because it would have saved him from some hate speech

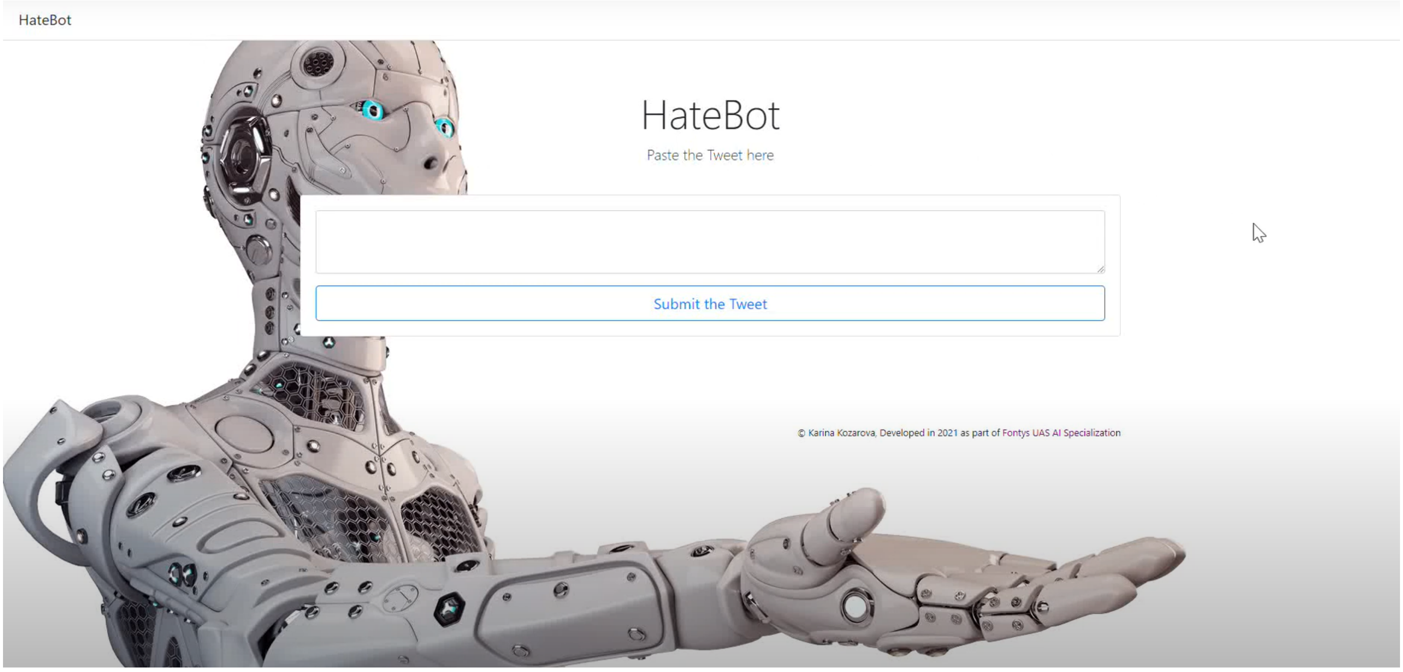
With the help of that machine learning model, we also make sure that:

- 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.

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’ which was the initial goal of the HateBot project.

Implementation result

To demonstrate the end result better, a web application was implemented using Django and Bootstrap. A short demo of that application can be viewed on YouTube. The source code of it is available on GitHub. In short, the application is a website where the user can enter the text of a Tweet and after submitting it receive an answer if it is hate speech or not.



© Karina Kozarova, Developed in 2021 as part of Fontys UAS AI Specialization

Also, an API is provided where the user can just specify the Tweet text and receive a boolean telling if it’s hate speech or not. This way, even more systems can be created.

All of the code (both for the web application and the API) is open sourced in hopes that it might help someone out there looking for a way to classify Tweets as hate speech or not.