

h8 identifier - no space for hate

Hateful posts and cyberbullying disturb internet users

Disturbing posts on minorities, political opponents and content creators can be found all over the internet. Some might be discomforting but others may be threatening. There are even structured, organized threats from right-wing trolls and conspiracy theorists, who deliberately torpedo discussions and flood the news feeds and comments of political opponents with hate, insults and death threats. In short: there should be no space for hate speech on the internet. The riots of the Capitol was just one of many events which could possibly be avoided if hate campaigns via social media were detected and the spread blocked.

The idea of the h8 identifier was born.

We wanted to train a classifier, which can differentiate between hate and non-hate speech. h8 identifier should give a warning if the threshold for too much hate is reached before such a message can be spread to cause more harm.

We started with a team of five people with different backgrounds but one common interest: motivation to make the internet a friendlier place. In our vision the h8 identifier should be displayed as a hate seismograph to screen hateful comments and posts. We focused on youtube, one of the most frequented social media platforms nowadays. Our target group is the youtuber, who can disable these comments, and also people who want to enjoy a video without hateful comments. The h8 identifier should work as a browser extension rather than a bot and ideally should have its own website for users to have a review and give feedback.

Our team consists of three more experienced and two newbie-yet-motivated techies. Our project divided into two chunks:

- A. Developing a Machine Learning pipeline which takes comments of a youtube video and generates a feasible prediction of the hate.
- B. Developing a website and visualize the personalized results of the hate in statistics relevant for the user

A The Machine Learning pipeline: from the youtube url to the hate prediction

1. Input Youtube comments in a JSON with a Youtube API.

Google offers multiple Youtube APIs for different usage. The official Python client library can be found here [googleapiclient](#). We found a suitable API to export youtube comments in a JSON file. Thanks to this helpful [repository](#).

One major limitation of the the Youtube API ([Comments: list](#) | [YouTube Data API](#) | [Google Developers](#)) is the limitation of 100 items per request. Nevertheless, this enables us to predict the hatefulness of the most recent 100 comments of a video.

Our solution for that problem: Using pageToken parameter from the API we can analyze all comments and subcomments. We successfully implemented that feature on 30.01.2021.

Next step is to filter for youtube comments ignoring irrelevant information such as the comment author's name.

2. Format and clean the data for the classifier to understand

We extracted the text of comments and subcomments. Then we organized them in a pandas dataframe for further analysis. Major concerns were non -text characters such as @, #, : , etc. The [tweet-preprocessor library](#) was a good start to clean the data.

Since our model won't understand human language, the cleaned comments are converted in a numerical form: First the comments are tokenized into words, then the model will go through the stack of comments to count for each word. For that task we use the CountVectorizer from sklearn.

3. Choose a suitable classifier, train and pickle it

We searched for a natural language processor and a binary classifier which is trained on an ideally large pre labeled dataset: either hateful text or non hateful text. First we found github repositories that specialized in hate classification for text. However most of them where not actively maintained or were 2 years old using python 2 which caused problems in package version compatibility. One of them used a binary classifier BERT provided by tensorflow, which was primarily interesting for us since it is still actively maintained. However it required a tensorflow 2.4 version to run whereas in the script was based on tensorflow 1 and consists of packages not compatible with tensorflow 2.4.

For the start we decided to use a simple NLP classifier, which worked well in the [Twitter Sentiment Analysis](#) and does not cause problems in version compatibility. We applied a supervised learning model, calles Support Vector Classifier (SVC), which can be used for binary or multi class classification onto our data.

We trained it with a large pre labeled data set of hateful vs. non-hateful twitter posts available on kaggle: <https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech>. Hateful comments are labeled as 1, the rest as 0. Using the train_test_split function of sklearn we divided the data set in 70% training and 30% test data. The accuracy of 95% convinced us to stay with this model and pickle it for further later application.

4. Create a function for the frontend

Having pickled the model and the vectorizer we defined a prediction() function to apply the model to data with at least one column with comments labeled as "tweet". The output is a list containing following elements:

- a pandas dataframe with the first 10 hateful (labeled as 1) comments
- the count of hate comments
- the count of all analyzed comments
- the ratio of hateful comments within all checked comments

For further analysis, we saved the following variables from our predict(df) function, which will not be displayed on our website due to the vast amount of comments:

- a numpy array called `y_pred_svm`: It contains the predictions for all the comments labeling 1 for hate and 0 for non-hate.
- `hateful_comments` is a list consisting of all comments (tweet) which are labeled as hate (1)

Then we integrated it to the request from Youtube API making a two new functions called `get_id_from_url(url)` and `get_predictions(videoID)`. The first function takes a youtube-url, parses the youtube video ID. The prediction function takes that ID for the get request and all comments are inputted into the pipeline returning the above mentioned predictions.

5. Possible Limitations of the model

First, we detected some errors in the labeling of the comments during our cross check of the training data. We then created a file with comments to test our model. Curiously comments with “so trash” were labeled wherase “shit” was not considered as hate. Moreover long comments are more likely to be labeled hateful than short comments.

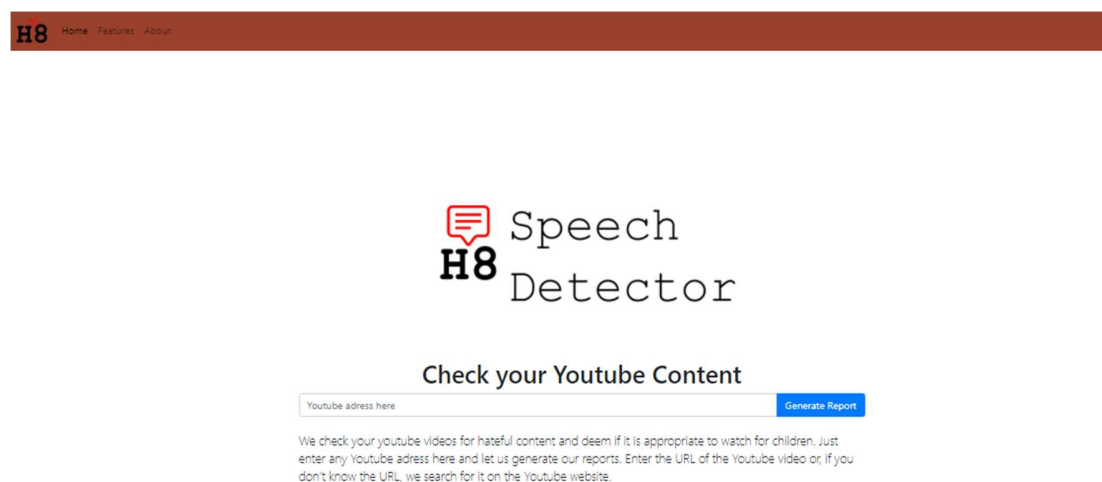
Second, nearly all of the comments were in English, thus comments in other language might be falsely labeled as hate. Therefore our model is only to be applied on English comments.

Nevertheless the majority of us decided to stick with this classifier to keep the pipeline going. In case we find a more suitable classifier in the future or more accurate training data, we can replace it.

B Our result

Using the flask framework and the UX Design drafted with figma we created a the h8 identifier web application.

Our homepage can also be accessed when selecting the H8 symbol.



In case no valid youtube url was inputted, this error message is shown.



Something went wrong...


Was that really a Youtube video-URL?

Retry

If a valid youtube video was inputted, we display the video details including the views, comments, likes and dislikes.

H8 Home Features About

Mark Morgan: 'Open borders' rhetoric is driving new migrant caravans



Fox News

Acting CBP Commissioner Mark Morgan said he believes the incoming Biden administration's immigration plan is fuelling new migrant caravans in Central America. Subscribe to Fox News! <https://bit.ly/2va...>

2912 1.89% Views: 207655

2562 304

We analyzed a representative number of comments calculate the hate ratio, give a feedback to the user according to the hate ratio that ratio. Then we show the 10 comments with highest hate probability below. A disclaimer is added "If they are not hateful to you, you can enjoy your youtube experience without worries."

This video has 207655 views which means that 1.4 out of 100 viewers have left a comment. We have analyzed the first 2379 comments. 1.89% of these comments are hateful, discriminatory or racist according to our algorithm.

H8 identifier detected a very low percentage of hate comments. It is unlikely that you will encounter a hateful comment during your youtube experience.

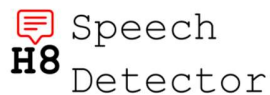
The 10 comments with highest hate probability are displayed below. If they are not hateful to you, you can enjoy your youtube experience without worries.

- Great Biden!! Good job you stupid pretend president..
- We need citizens with guns to do the governments job.
- @Lee Harrison Can you tell me who's worse? Trump and his Congressional minions or his brainless treasonous followers?
- The Biden administration has no love for this country. Democrats are willing to sell America one plot at a time. They aren't willing to help Americans who are in deep financial straits due to their lockdown policies and will take them a very long time to put America back on its feet. His very first statement was to increase taxes. He will overburden it's citizens with policies that will curtail investments and production. His policies are already gearing to welcome back CCP that put America at risk. They were benefiting from their largess and woe to downtrodden Americans. CCP connections are making them rich to hell with the rest of the citizenry. I just wish they should all join their friends the CCP in China and leave the country alone. There are rich citizens also who hate Trump and they come from Europe but are American citizens. They consider Trump a threat because of his popularity. They don't want the status quo to change. China is not our only enemy. These people are shadow funding organizations that would compromise the country and couldn't care less for the country as long as they're making money. These are people who stay here only a few months at a time and are usually in Germany or Switzerland.
- Cornell University research found liar trump is the biggest driver of coronavirus disinformation.
- no boundaries, no borders, riots all summer, unaccountable abuse of an Elected President, for Years, unemployed Americans, and now 20,000 troops in Nation's Capitol? - makes as much sense as socialism these people
- So Biden us telling them not to come, double talk, talking out if the other side of his mouth. See Democrat's he told you what you wanted to hear. When he was campaigning. How about the major tax brake to the billionaires
- Trump is migrating south-to Guantanamo.
- The open border rhetoric is coming from Republicans not Biden. In fact, Biden is working with Guatemala and Mexico to stop the caravans. If Fox wants to help the country, it should report the news not lies.

The about page gives more insights about our project goal.

H8 [Home](#) [Features](#) [About](#)

No space for hate



Hateful posts and cyberbullying disturb internet users

Disturbing posts on minorities, political opponents and content creators can be found all over the internet. Some might be discomforting but others may be threatening. There are even structured, organized threat from right-wing trolls and conspiracy theorists, who deliberately torpedo discussions and flood the news feeds and comments of political opponents with hate, insults and death threats. In short: there should be no space for hate speech on the internet. The riots of the Capitol was just one of many events which could possibly be avoided if hate campaigns via social media were detected and the spread blocked.

The idea of the h8 identifier was born

We wanted to train a classifier, which can differentiate between hate and non-hate speech. h8 identifier should give a warning if the threshold for too much hate is reached before such a message can be spread to cause more harm. We started with a team of five people with different backgrounds but one common interest: motivation to make the internet a friendlier place. In our vision the h8 identifier should be displayed as a hate seismograph to screen hateful comments and posts. We focused on youtube, one of the most frequented social media platforms nowadays. Our target group is the youtuber, who can disable these comments, and also people who want to enjoy a video without hateful comments. The h8 identifier should work as a browser extension rather than a bot and ideally should have its own website for users to have a review and give feedback.


Contributors:

- Urs Schmidt
- Thuy Anh Nguyen
- Robin van de Water UX

We thank our mentor Felix Linker who kept us motivated and gave us guidance and input whenever needed.

Thank you Maiuran and Wahid for your input.

References

1 Youtube API [Comments: list](#) | [YouTube Data API](#) | [Google Developers googleapis/google-api-python-client](#):  The official Python client library for Google's discovery based APIs. ([github.com](#))

2 NLP classifier [Twitter-Sentiment-Analysis/Twitter Sentiment Analysis Support Vector Classifier.ipynb at master · importdata/Twitter-Sentiment-Analysis \(github.com\)](#)