Overview: The exponential growth of social media has brought with it an increasing propagation of hate speech. The ease of communication nowadays has led to an immense increase of exchange of ideas between people, but what was supposed to be a positive dynamic led to one of the biggest dualities of the moment: how to navigate between freedom of speech and the threatening use of insulting and offensive language? To bring the matter into closer context, we often observe YouTube/Twitch administrators frequently facing themselves with the task of manually banning users from their channels.

Countries are currently engaging in initiatives aiming to develop effective counter-measures against this practice and one of the first challenges involved in this combat is identifying hate speech via automatic processes that enable effective decision-making from authorities. Thus, in this project, we aim to build a hate speech detection model that classifies a dataset of tweets into hate or non-hate comments.

Data:

The dataset we are using compromised of ~25k tweets, labeled as hate speech, offensive, or neither. Using the Twitter API, tweets that contained words and phrases identified by internet users as hate speech (compiled by Hatebase.org) were queried, and the users behind these tweets were selected. All of those users' lifetime tweets were then extracted, resulting in a corpus of 85.4 million tweets. From this corpus, a random sample of 25k tweets that contained hate speech words and phrases were selected. These tweets were then hand labeled by three or more people, who were provided with a definition of hate speech (language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group). In case of disagreement between annotators, majority opinion ruled (the overall agreement score between annotators was 92%). Tweets that did not have a majority opinion were discarded. The dataset is the resulting corpus of 24,802 labeled tweets. In the dataset, 5% of the tweets were labeled as hate speech, 76% as offensive, and the remainder as neither.

Method: We will first load the data into a pandas dataframe, splitting it up into an 80-20 train-test split. Next we will train several models on the dataset using different classifiers such as Bert, K-NN, logistic regression, and naive bayes in order to find which model selects the best

features and predicts each of the 3 classes most accurately. We will then use 10 fold cross validation in order to ensure that the model generalizes well and then proceed to test it against the remaining 20% of data.

Evaluation: To evaluate the models we will be using, we plan to test performance based upon accuracy, recall and the averaged F1 score. Given that there isn't a metric within the space that has been proven to be the most appropriate, we believe that collectively comparing the accuracy, recall, and F1 score will provide us with a good basis as to which model performs the most optimally. In addition, given the imbalance between the number of hate speech and non-hate speech tweets within the dataset, solely looking at accuracy will not suffice in finding the best model (a high accuracy may not be as difficult to attain given the imbalance).

References:

Tontodimamma, A., Nissi, E., Sarra, A. et al. "Thirty years of research into hate speech: topics of interest and their evolution". *Scientometrics* 126, 157–179 (2021).

Thomas D, Dana W, Michael M, Ingmar W "Automated Hate Speech Detection and the Problem of Offensive Language"

Zeerak Waseem and Dirk Hovy. 2016. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The Effects of User Features on Twitter Hate Speech Detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.