Automated Hate Speech Detection and the Problem of Offensive Language

COMP8240 S2 2023 - Group Project

Chandrashekhar Meenakshisundaram
Jerome Roosan
Raja Pedapudi
Sri Venkata Kameswara Naga Phanindra Kavipurapu

Macquarie University

October 9, 2023

Contents

- Recap of paper
- Recap of data
- Recap of code
- Progress of the project
- Construction of new datasets

Recap of paper

- Offensive speech is a prevalent problem in online spaces. This makes automated hate speech detection a necessity.
- The current paper distinguishes between offensive and hateful speech whereas others usually don't. It uses a crowd-sourced hate speech lexicon to collect tweets with hateful keywords.
- These are classified as hate speech, offensive speech, and neutral speech. Multiclass classification is used to distinguish between the three types.
- In some cases, it is easier to differentiate between hate speech and offensive speech while it is not as easy in others. Tweets that have racist or homophobic keywords are readily classified as hate speech whereas sexist keywords lead to classification as offensive speech. The absence of explicit keywords makes it harder to classify as anything.

Recap of data

- Twitter API is used to obtain tweets (85 million in total) that have keywords present in a hate speech lexicon.
- 25,000 of them were randomly sampled and annotated manually and labeled as hateful, offensive, or neutral.
- Each tweet was coded by 3 or more people and the majority vote served as the final label. The context in which the words were used was considered for better performance.
- 5% of tweets were coded as hate speech, 76% were considered as offensive and the rest as neutral.

Recap of code

- Pandas, NumPy, scikit-learn, and NLTK are used to build the classifier.
- Data is first preprocessed and then tokenized using Porter stemming. Tweets are taken as a list of strings on which POS tagging is performed.
- Additional features are obtained like sentiment scores, readability scores, etc. Term frequency and inverse document frequency vectorizations are performed.
- Logistic regression is used as metatransformer to obtain the important variables which are then fed to a support vector classifier.
- The model reports an average precision and recall of 0.91 and 0.90 and an F-1 score of 0.91.

Progress of the project

- A VM has been created with *Ubuntu 22.4*
- The project is sourced from GitHub using **git** installed in the VM
- Jupyter Notebook is being used to execute the Python Code.
- The dataset available with the paper and code are being validated.

Construction of new datasets

- Paper works with a dataset of tweets by Donald Trump, retrieved from Twitter.
- Since **X** API is currently paywalled, alternatives were explored.
- The main option that we are exploring is scraping the Internet for **Threads** since there is no official API
- Another possibility is to use existing datasets containing tweets that were flagged for offensive content and hate speech.
- Offensive and hate speech Tweets datasets have been identified in Kaggle.