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

August 28, 2023

Contents

Abstract of paper

Abstract of paper

- There are many types of offensive speech prevalent on the internet. The main challenge here is to differentiate between what is plain offensive and what is hateful.
- Current methods make no distinction between offensive and hateful speech.
- Lexical detection methods and unsupervised learning methods have failed to differentiate between them.

Abstract of paper

- This paper uses crowd-sourced hate speech lexicon to collect tweets that have keywords related to offensive and hateful speech.
- CrowdFlower, an AI company that uses humans to label data, is used to label the samples into 3 categories - hate speech, offensive language, and neutral speech.
- Multi-class classification is used to distinguish between the 3 types of speech listed.

Abstract of paper

- A close analysis of the results shows that in some cases, it is easy to differentiate between hate speech and offensive speech but in other cases it is harder.
- Tweets with words that are deemed racist and homophobic are more likely to be classified as hate speech.
- Tweets with sexist keywords are likely to be classified as offensive speech. Tweets that don't have explicit keywords are more difficult to classify.

Abstract of code

Abstract of code

- The main packages for building the classifier are pandas, NumPy, scikit-learn, and NLTK.
- It starts with the preprocessing of the data where the URLs are all replaced by a single string ‘URLHERE’.
- Multiple whitespaces are replaced by a single whitespace, and all the mentions are replaced by ‘MENTIONHERE’. Regular expressions are used for this purpose.

Abstract of code

- Then the data is tokenized using Porter Stemmer technique.
- Stopwords are removed as well for cleaning.
- The tweets are taken as a list of strings on which parts-of-speech tagging is performed.
- The urls, mentions, and hashtags are counted.

Abstract of code

- A key component of preprocessing is obtaining additional features.
- There is a function that takes a string and returns some additional features such as sentiment scores, text and readability scores, number of words, average syllables, number of unique terms, etc.
- Finally, tweets are converted into a format that can be used as input to the model using term-frequency vectorizer, inverse document frequency vectorizer.

Abstract of code

- In summary, each tweet is decomposed to obtain an array of TF-IDF scores for a set of n-grams in the tweets, an array of parts-of-speech tags in the tweets, and an array of features including sentiment readability, and vocabulary.
- Logistic regression is used as a meta-transformer to obtain the important variables which are then fed to a support vector classifier.
- The model reports an average precision of 0.91, recall of 0.90, and an F1 score of 0.91.

Abstract of data

Abstract of data

- Using the Twitter API, tweets containing words present in a hate speech lexicon (compiled by Hatebase.org) were identified, resulting in a set of 85.4 million tweets.
- A random sample of 25,000 tweets were obtained, which were manually labeled by CrowdFlower. The labels are neutral speech, offensive speech, and hate speech.
- Each tweet was coded by three or more people and the majority vote was taken as the label. The inter-coder agreement provided by CrowdFlower is 92

Abstract of data

- For labeling, specific instructions were provided. The raters were asked to consider the context in which the words were used.
- The presence of a word, no matter how offensive, did not necessarily indicate a tweet is hateful or offensive.
- This resulted in a sample of 24,802 labeled tweets.
- Given the stricter criteria for hate speech, only 5% of tweets were coded as hate speech. The majority (76%) of the tweets were considered offensive, and the rest were considered neutral or non-offensive.

Previously considered papers

Previously considered papers

- A Multi-source Graph Representation of the Movie Domain for Recommendation Dialogues Analysis[1]
- Suggest me a movie for tonight: Leveraging Knowledge Graphs for Conversational Recommendation[2]

[1] https://aclanthology.org/2022.lrec-1.138.pdf

[2] https://aclanthology.org/2020.coling-main.369.pdf

Source & justification

Source & justification

- Presented in **International Conference on Web and Social Media (ICWSM)** [3]
- ICWSM is **ranked 11 in Google Scholar's Top Publications for Databases and Information Systems**
- ICWSM's h-5 index is **59**
- ICWSM's h-5 median is **82**
- **712** Github stars, **147** Citations and **15** references on **https://paperswithcode.com**[4]

[3] Source: Databases & Information Systems - Google Scholar Metrics. https://scholar.google.com.au/citations?view_op=top_venues&hl=en&vq=eng_databasesinformationsystems

[4] https://paperswithcode.com/paper/automated-hate-speech-detection-and-the