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 30, 2023

Contents

Recap of paper

- The current paper distinguishes between offensive and hateful speech. The original paper used 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.
- 25,000 of them were randomly sampled and annotated manually and labeled as hateful, offensive, or neutral. The dataset given in the source work had 5% of tweets which were coded as hate speech, 76% considered as offensive and the rest as neutral.
- **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.

Replication of the source work

- The authors provided the necessary code and data via their GitHub repository. The repository had a notebook that we executed to replicate the original results.
- We had perfect replication of the source work. We did not have to subset the data as it only had 25,000 records. We could execute the code on our local machines. We successfully replicated the original work.
- Our main issue was being stuck in dependency hell due to the outdated versions of Python (2.7) and Python packages. Many functions had been deprecated and attributes had changed causing many errors.

Output of the source work

	precision	recall	f1-score	support
0	0.24	0.54	0.33	164
1	0.93	0.77	0.84	1905
2	0.60	0.78	0.68	410
accuracy			0.76	2479
macro avg	0.59	0.70	0.62	2479
weighted avg	0.83	0.76	0.78	2479

Figure. 1: Precision, Recall & F1-Score of Source Work

Output of the source work

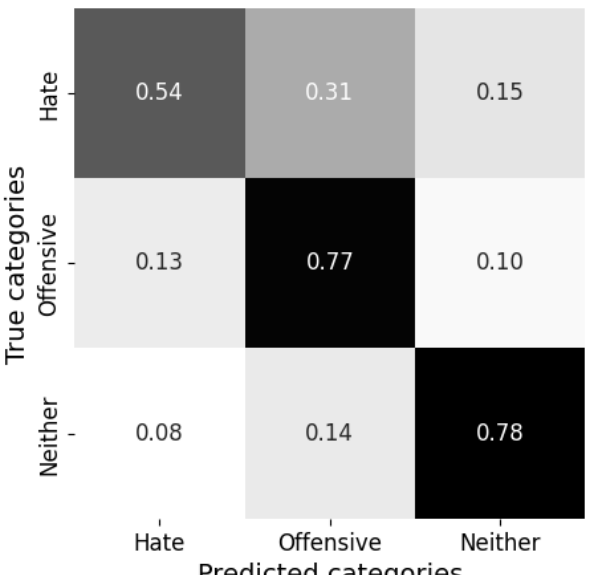

Figure. 2: Confusion Matrix of Source Work

Construction of new dataset

- A new dataset of 30,000 tweets was obtained from Kaggle.
- The downloaded data came with labels that were incompatible with our model, and as we couldn't pay for labeling, we subset 1000 tweets from the 30,000 and labeled them by hand. Each of us labeled 250 tweets.

Evaluation of new data set

- The source notebook had methods for evaluation that we used.
- It involved obtaining precision, recall, and F1 scores for the model.
- The F1 score is important due to the imbalance in the dataset.

OUTPUT OF NEW DATASET

	precision	recall	f1-score	support
0	0.29	0.50	0.36	4
1	0.82	0.86	0.84	64
2	0.77	0.62	0.69	32
accuracy			0.77	100
macro avg	0.63	0.66	0.63	100
weighted avg	0.78	0.77	0.77	100

Figure. 3: Precision, Recall & F1-Score of New Data

Output of new dataset

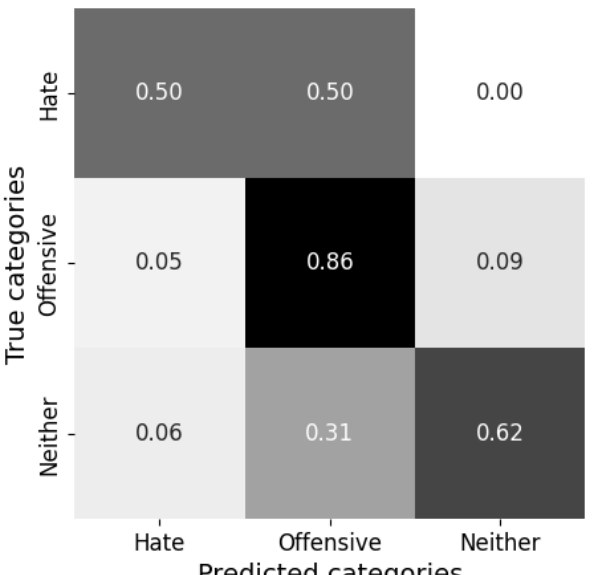

Figure. 4: Confusion Matrix of New Data