Implementation of Automatic Detection of Hate Speech in Text

Kunal Panjwani
MSc. Data Science
DAIICT
Gandhinagar
202018001@daiict.ac.in

Vanditha Vinod
MSc. Data Science
DAIICT
Gandhinagar
202018003@daiict.ac.in

Nihar Shah
MSc. Data Science
DAIICT
Gandhinagar
202018014@daiict.ac.in

Aakanksha Shah
MSc. Data Science
DAIICT
Gandhinagar
202018026@daiict.ac.in

Yagn Purohit
MSc. Data Science
DAIICT
Gandhinagar
202018035@daiict.ac.in

***Abstract*—In this project, we have created a model to automatically detect & classify text into hate speech & offensive language. A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. Previous work using supervised learning has failed to distinguish between the two categories. Crowd-sourcing is used to label a sample of tweets into three categories: those containing hate speech, those containing offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories.**

***Index Terms*—Hate Speech, Text Analysis, Natural Language Processing (NLP), Sentiment Analysis, Machine Learning**

I. Introduction

In recent times, the rise of hate speech on social media platforms has increased exponentially with the increasing number of internet users. People have the freedom to speech and can anonymously comment on whatever and however they want. This power is being abused and is creating a negative impact on many other people psychologically.

What constitutes hate speech and when does it differ from offensive language? No formal definition exists but there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them [1]. In many countries, there are laws prohibiting hate speech, which tends to be defined as speech that targets minority groups in a way that could promote violence or social disorder. People convicted of using hate speech can often face large fines and even imprisonment. These laws extend to the internet and social media, leading many sites to cre- ate their own provisions against hate speech. Both Facebook and Twitter have responded to criticism for not doing enough to prevent hate speech on their sites by instituting policies to prohibit the use of their platforms for attacks on people based on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence towards others.

Drawing upon these definitions, according to ”A Survey on Automatic Detection of Hate Speech in Text” by Paula Fortuna & Sergio Nunes [2], hate speech is defined as ”Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.”

Importantly,we are not including all instances of offensive language because people often use terms that are highly offensive to certain groups but in a qualitatively different manner. For example people use offensive terms when quoting rap lyrics, even when they don't intend to offend anyone in particular. Such language is prevalent on social media, making this boundary condition crucial for any usable hate speech detection system.

II. Data

The authors of ”Automated Hate Speech Detection and the Problem of Offensive Language” [3] began with a hate speech lexicon containing words and phrases identified by internet users as hate speech, compiled by Hatebase.org [4]. Using the Twitter API they collected tweets containing terms from the lexicon, resulting in a sample of tweets from 33,458 Twitter users. They extracted the time-line for each user, resulting in a set of 85.4 million tweets. From this corpus they then took a random sample of 25k tweets containing terms from the lexicon and had them manually coded by CrowdFlower(CF) workers. The workers were asked to label each tweet as one of three categories:

- Hate speech
- Offensive but not hate speech
- Neither offensive nor hate speech

Users were asked to think not just about the words appearing in a given tweet but about the context in which they were used. They were instructed that the presence of a particular word, however offensive, did not necessarily indicate a tweet is hate speech. Each tweet was coded by three or more people. The intercoder-agreement score provided by CF is 92%. They used the majority decision for each tweet to assign a label. Some tweets were not assigned labels as there was no majority class. This results in a sample of 24,783 labeled tweets. Only 5% of tweets were coded as hate speech by the majority of coders and only 1.3% were coded unanimously, demonstrating the imprecision of the Hatebase lexicon. This is much lower than a comparable study using Twitter, where 11.6% of tweets were flagged as hate speech [5], likely because they used a stricter criteria for hate speech. The majority of the tweets

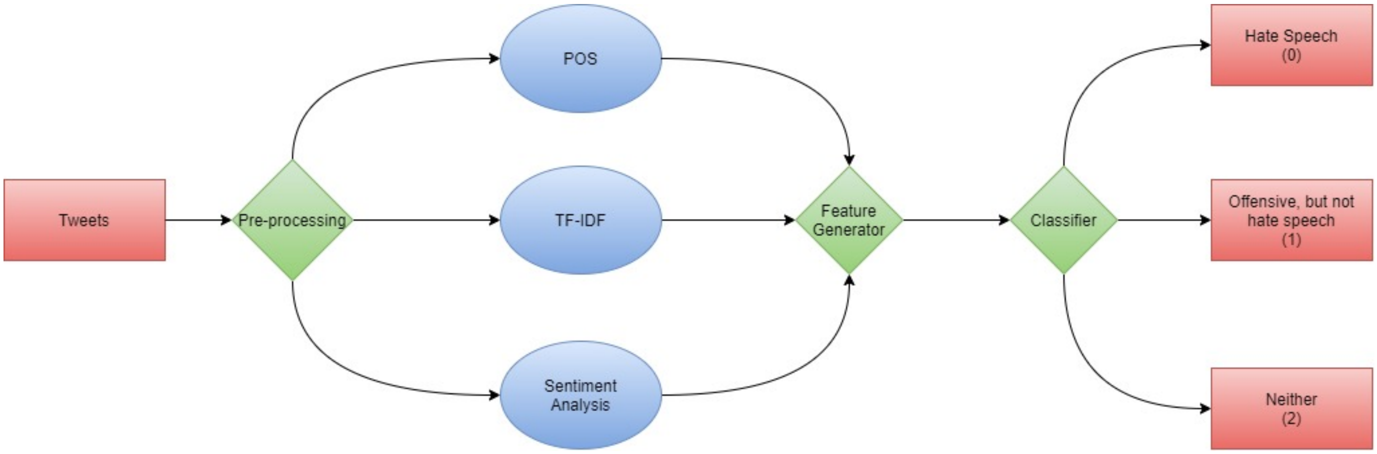


Fig. 1. Flow of the entire project.

were considered to be offensive language (76% at 2/3, 53% at 3/3) and the remainder were considered to be non-offensive (16.6% at 2/3, 11.8% at 3/3). They then constructed features from these tweets and used them to train a classifier.

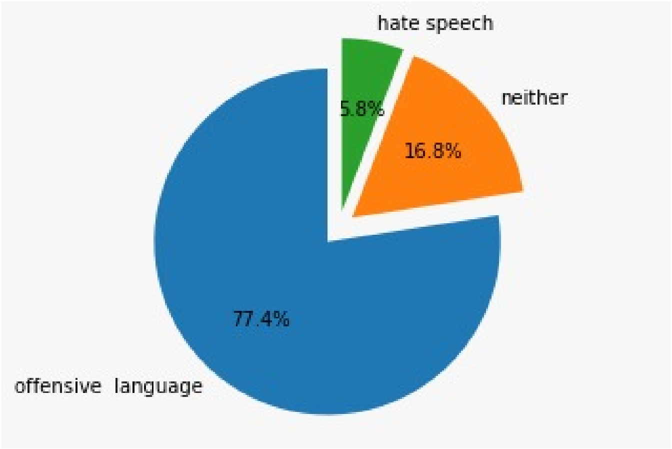


Fig. 2. Pie chart of label comparison.

III. Features

We lowercased each tweet and tokenised using RegexpTokenizer, and weighted each word by its TF-IDF. To capture information about the syntactic structure, we used NLTK [6]to construct Part-of-Speech (POS) tag, unigrams, bigrams, and trigrams. To capture the quality of each tweet we use modified Flesch-Kincaid Grade Level and Flesch Reading Ease scores, where the number of sentences is fixed at one. We also use a sentiment lexicon designed for social media to assign sentiment scores to each tweet [7]. We also included binary and count indicators for hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet.

IV. Model

We first use a logistic regression with L1 regularization to reduce the dimensionality of the data. We then test a variety of models that have been used in prior work: logistic regression, naıve Bayes, decision trees, random forests, and linear SVMs. We tested each model using 5-fold cross validation, holding out 10% of the sample for evaluation to help prevent overfitting. After using a grid-search to iterate over the models and parameters we find that the Logistic Regression tended to perform significantly better than other models. We decided to use a logistic regression with L2 regularization for the final model as it more readily allows us to examine the predicted probabilities of class membership and has performed well in previous papers [5] [8]. We trained the final model using the entire dataset and used it to predict the label for each tweet. We use a one-versus-rest framework where a separate classifier is trained for each class and the class label with the highest predicted probability across all classifiers is assigned to each tweet. All modeling was performing using scikit-learn [9].

V. Results

We applied various classification methods like Gaussian Naive Bayes, Linear SVC, and Random Forest Classifier, but Logistic Regression with L2 regularization gave us the best results. We applied 5-fold cross-validation on logistic regression and obtained the results as shown in Fig. 3.

As we can see, 69% of the tweets were correctly classified as ”Hate Speech”, which shows that the model is capable of distinguishing ”Hate Speech” & ”Offensive Language”.

VI. Conclusion

Given the legal repercussions of hate speech, it is important to correctly differentiate between commonplace offensive language & serious hate speech. Close analysis of the predictions and the errors shows when we can reliably separate hate speech from other offensive language and when this differentiation is more difficult. We find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Tweets without explicit hate keywords are also more difficult to classify.

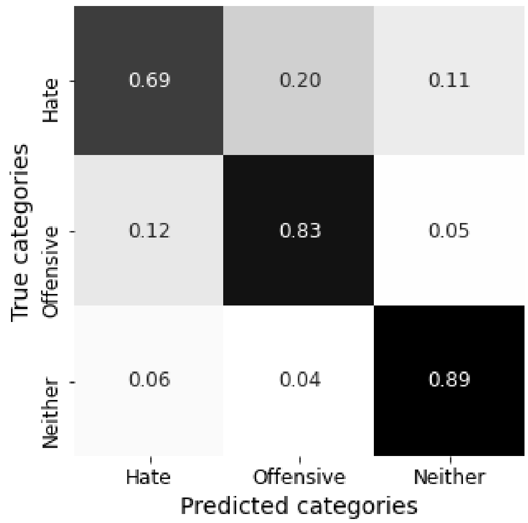


Fig. 3. Confusion Matrix to visualize the accuracy of the model.

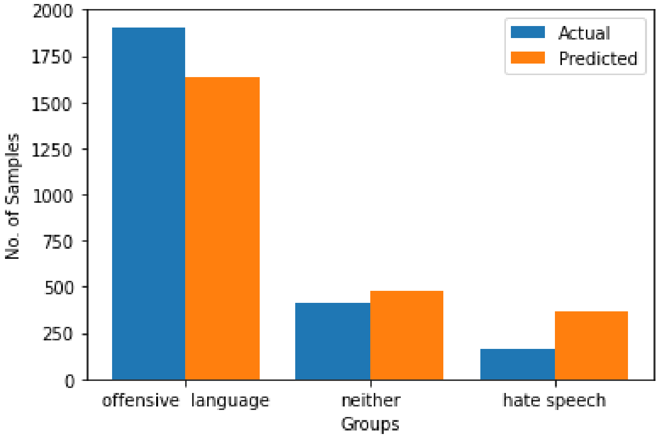
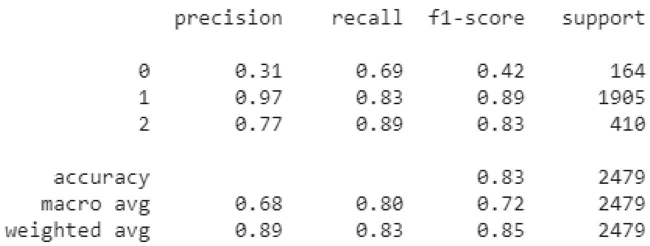


Fig. 4. Bar chart of train & test data comparison.



```
              precision    recall  f1-score   support

           0       0.31      0.69      0.42       164
           1       0.97      0.83      0.89      1905
           2       0.77      0.89      0.83       410

    accuracy                           0.83      2479
   macro avg       0.68      0.80      0.72      2479
weighted avg       0.89      0.83      0.85      2479
```

Fig. 5. Accuracy, Precision, & F1-Score.

VII. Contribution of members

All the members have been present at every meet and each one of us gave his/her input at every stage of the project. We have distributed the contribution on the basis of who gave the maximum input.

- Pre-processing - Nihar Shah (202018014)
- POS, TF-IDF, Sentiment Analysis & Feature Generation - Kunal Panjwani (202018001) & Aakanksha Shah (202018026)
- Classification - Vanditha Vinod (202018003) & Yagn Purohit (202018035)
- Report - Vanditha Vinod (202018003) & Aakanksha Shah (202018026)

References

[1] A. Rosga, "Samuel walker, hate speech: The history of an american controversy, lincoln, nebraska and london: University of nebraska press, 1994. pp. ix, 217. $11.95 (isbn 0-8032-9751-3)." *Law and History Review*, vol. 14, no. 1, p. 191–193, 1996.

[2] Fortuna, Paula and Nunes, Sérgio, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Comput. Surv.*, vol. 51, no. 4, July 2018. [Online]. Available: https://doi.org/10.1145/3232676

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, May 2017. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14955

[4] Hatebase. Hatebase is a service built to help organizations and online communities detect, monitor and analyze hate speech. [Online]. Available: https://hatebase.org/

[5] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85

[6] E. Loper and S. Bird, "Nltk: The natural language toolkit," *CoRR*, vol. cs.CL/0205028, 2002. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr0205.htmlcs-CL-0205028

[7] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text." in *ICWSM*, E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, Eds. The AAAI Press, 2014. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.htmlHuttoG14

[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93. [Online]. Available: https://www.aclweb.org/anthology/N16-2013

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.