Tejas Padmanabhan, Fall 2024

Analyzing the duality of comments on a Kaggle Dataset to detect hate speech(1478 words)

Research Question

What is the effectiveness of a supervised learning model in accurately detecting hate speech in a predetermined dataset of comments?

Introduction

Detecting hate speech when it comes to social media has always been a task that many platforms have tried to solve and erase. Supervised learning has been an aspect of machine learning that has been on the rise across the world as Artificial Intelligence becomes prevalent in all areas of our life. By using supervised learning, it becomes easier for social media platforms to detect hate speech in comment sections, posts, and more. According to Hameda Sennary, Most of the hate speech detection algorithms that exist today are dependent on supervised machine learning algorithms (Sennary). Sennary continues to say that the reason this is the case is due to the fact that the model is able to find a link between characteristics and predictions by using training data (Sennary). “A model’s ability to achieve this goal is measured by its generalization performance.” (Sennary). And this ability is what Sennary says is used to allow platforms to detect hate speech. However, according to Pyingkodi, supervised machine learning does have “limitations and can be prone to errors, particularly if the training data is biased or if the algorithms are not carefully designed and evaluated.” (Pyingkodi). He goes on to say that we must carefully consider the consequences of machine learning and that the methods used must be fair in order to take out biases/errors (Pyingkodi).

Data

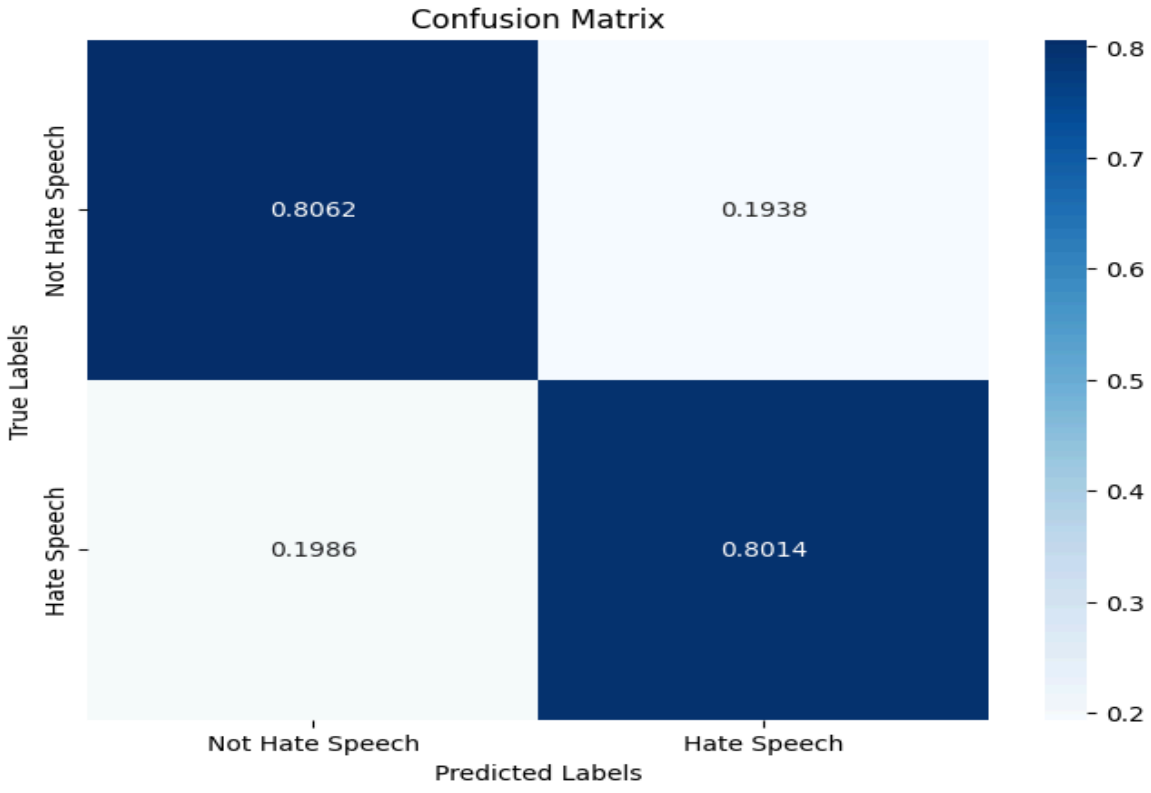
The dataset that I will be using is a pre-existing dataset from Kaggle. (https://www.kaggle.com/datasets/waalbannyantudre/hate-speech-detection-curated-dataset/data?select=HateSpeechDatasetBalanced.csv) This dataset contains over 726,000 data points that I used to train and create a model to detect hate speech. The timeframe for this dataset is unknown, however the dataset was published in 2022 which gives the indication that the data was collected around this year. I believe that this dataset is an appropriate dataset because it has a multitude of data points which makes it a great dataset. This dataset contains 2 columns, the actual content of the comments and a label, which has the value of either 0 or 1. If the value is 0, this means that the content of the comment is not hate speech and if the value is 1, vice versa, meaning the content does include hate speech. The actual content of the comments is from a variety of social media platforms including Twitter, Reddit, and other social media platforms. This dataset in terms of the specifics of the comments contains emoticons, emojis, hashtags, slang, and contractions required to detect hate speech on social media based on current trends.

Analysis

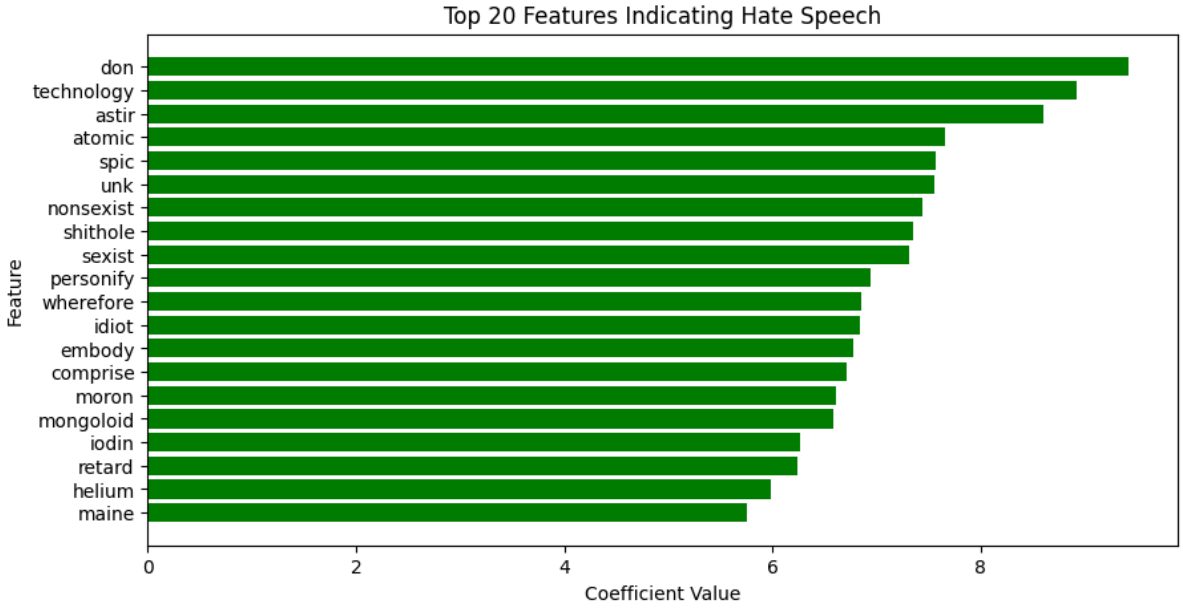
As exemplified before in the introduction, I am taking this dataset and completing both a supervised learning and unsupervised learning method to detect for hate speech. For the supervised learning method, I have taken the data and trained it and then proceeded to test it using the data that I gathered. I will create the code for the supervised learning method using Scikit-learn's TF-IDF model in order to detect hate speech. This will allow us to find the accuracy of the model in predicting hate speech. The coding for this method will be done using python and I will gather information I previously knew about training data to complete this method. I am using TF-IDF as the model because I have used TF-IDF before and understand that

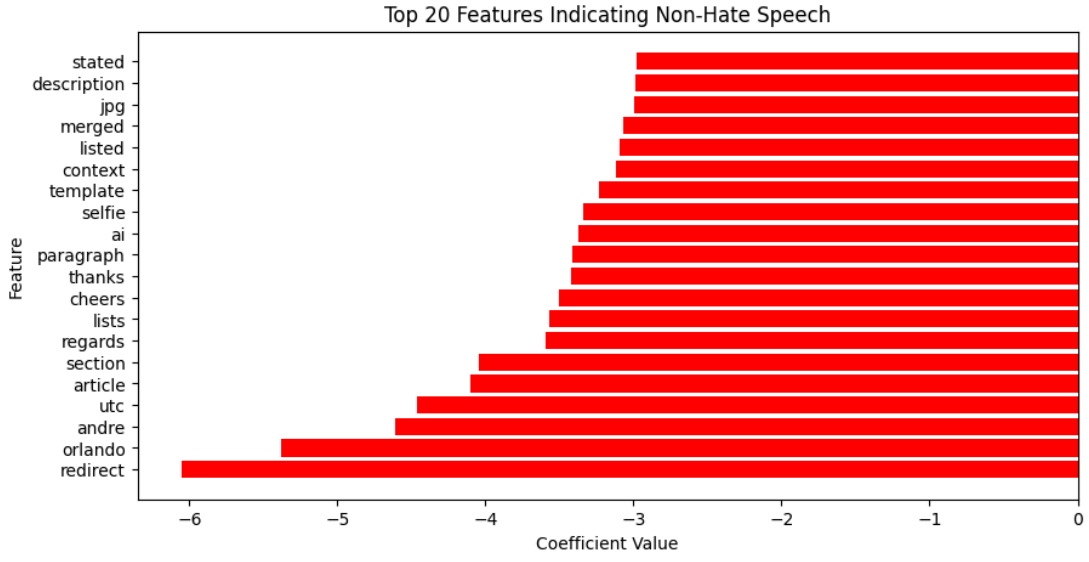
it has a unique document-level feature extraction capability which may make it better for detecting hate speech which we would want for our models. The results of the model will show us the accuracy, precision, recall, and F-1 score which are all important metrics for our model. One important note is that the test size I will be using is 0.2 to make sure that there is some variance for error otherwise the model may overfit and that is not good at all. As for the unsupervised learning method, I will be using topic modeling. In terms of specifics, I will be using an LDA as the codebase model and split words into different topics. This will allow us to see what kind of categories are created. I plan on creating 5 different kinds of categories and expect to see a split of categories showing hate speech and different categories with non-hate speech. For both the supervised and unsupervised methods, I created visualizations that both highlighted important points and showed results that show if our model was accurate/precise and categorizes words into different categories.

Results

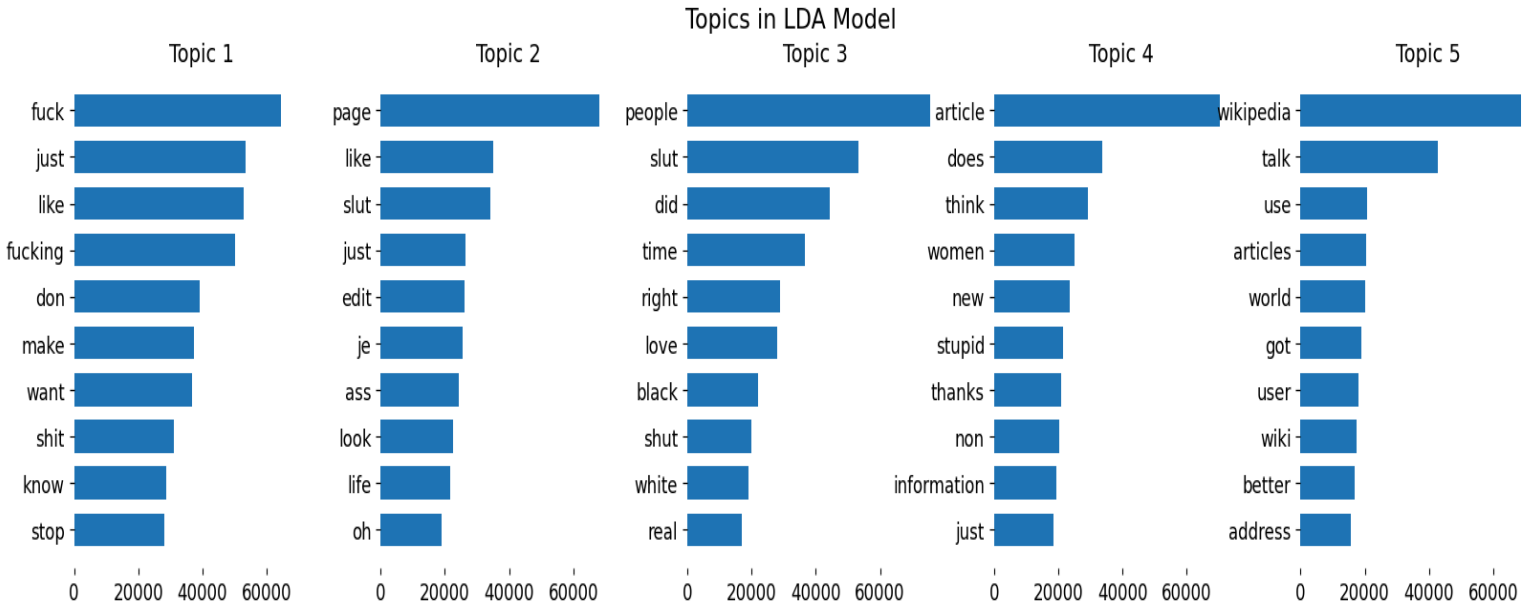


The confusion matrix using the TF-IDF supervised method helps us see how well the model tells the difference between hate speech and non-hate speech comments. The top-left box shows that about 80.6% of comments that are not hate speech were correctly identified, while 19.4% were wrongly marked as hate speech. The bottom-right box shows that about 80.1% of actual hate speech comments were correctly labeled, but 19.8% were incorrectly marked as not hate speech. Overall, the model does a good job of identifying most comments correctly, but there are still mistakes. Some normal comments are wrongly flagged as hate speech (false positives), and some hateful comments slip through as non-hate speech (false negatives). This means there's still room to make the model even better at avoiding these errors.





The two charts show the words that the model uses to decide whether a comment contains hate speech or not. In the first chart, we see the top words that indicate hate speech, such as "technology", "don", and "atomic", which the model found frequently in hateful comments. Some words, like "spic", "shithole", and "sexist" are clearly offensive, so the model strongly associates them with hate speech. In contrast, the second chart shows words that are linked to non-hate speech, such as "description," "jpg", "context", and "thanks". These words are more neutral or polite and are often found in normal, non-offensive comments. By analyzing these patterns, the model learns to recognize words that show negativity versus words that have neutral or positive content, helping it accurately identify hate speech. In the future if we run another dataset across the model, there is a very good chance the model recognizes the associations to these words and is able to correctly recognize the hate/non-hate speech.



This chart shows the results of the topic modeling which is the unsupervised learning method I used for this dataset. The chart displays five topics labeled Topic 1 through Topic 5 and each topic contains the most common words associated with it. For example, Topic 1 includes

words like "fuck", "just", "like", and "fucking" which means that the category relates to aggressive or explicit language. Topic 2 has terms like "page," "slut", "edit", and "like", possibly indicating online insults. Topic 3 has words such as "people", "slut", "did", and "love" again displaying derogatory language about individuals similarly to Topic 2. Topic 4 includes "article", "does", "women", and "stupid" pointing to gender-related discussions or different kinds of opinions online. Topic 5 has words such as "wikipedia", "talk", "use", and "world" which might relate to arguments or discussions occurring on the social media platforms mentioned in the introduction. Just based off of the five topics, I am surprised that the topic modeling didn't split the five topics into categories with hate speech and non-hate speech and instead split by topics we might see online. I would say topic 4 and 5 definitely had less hate speech than topic 1,2, and 3 which were the complete opposite. Overall, this chart helps identify patterns and themes in the hate speech dataset which were not previously seen before.

Conclusion

This project looked at how well a supervised learning model (TF-IDF) can detect hate speech in social media comments. The results showed that the model accurately predicted about 80% of the time for both hate speech and normal comments. However, it still made mistakes and some regular comments were wrongly labeled as hate speech, and a few hate speech slipped through as normal. When we looked at the words the model used to make decisions, we saw that offensive words were linked to hate speech, while polite or neutral words were connected to non-hate speech. I also ran an unsupervised learning method, topic modeling which grouped the comments into five topics, some filled with rude or aggressive language and others showing more normal discussions. This shows that while the TF-IDF works pretty well, it still needs

improvements, and the LDA topic modeling helps us find patterns in the comments. One issue with this project is that only one method was used for supervised learning and unsupervised learning. Also, we don't know exactly when or how the comments in the dataset were collected, which could change the results if we tried this again with different data.

Works Cited

Pyingkodi, M. "Hate Speech Analysis Using Supervised Machine Learning Techniques | IEEE Conference Publication | IEEE Xplore." *IEEE Xplore*, 24 May 2023, ieeexplore.ieee.org/document/10128591.

Sennary, Hameda A., et al. "Detection of Hate: Speech Tweets Based on Convolutional Neural Network and Machine Learning Algorithms." *Nature News*, Nature Publishing Group, 21 Nov. 2024, www.nature.com/articles/s41598-024-76632-2#:~:text=Most%2C%20if%20not%20all%2C%20of,to%20previously%20unseen%20inputs20.