MACHINE LEARNING PROJECT

Aamod B K
Ishaan Jalan

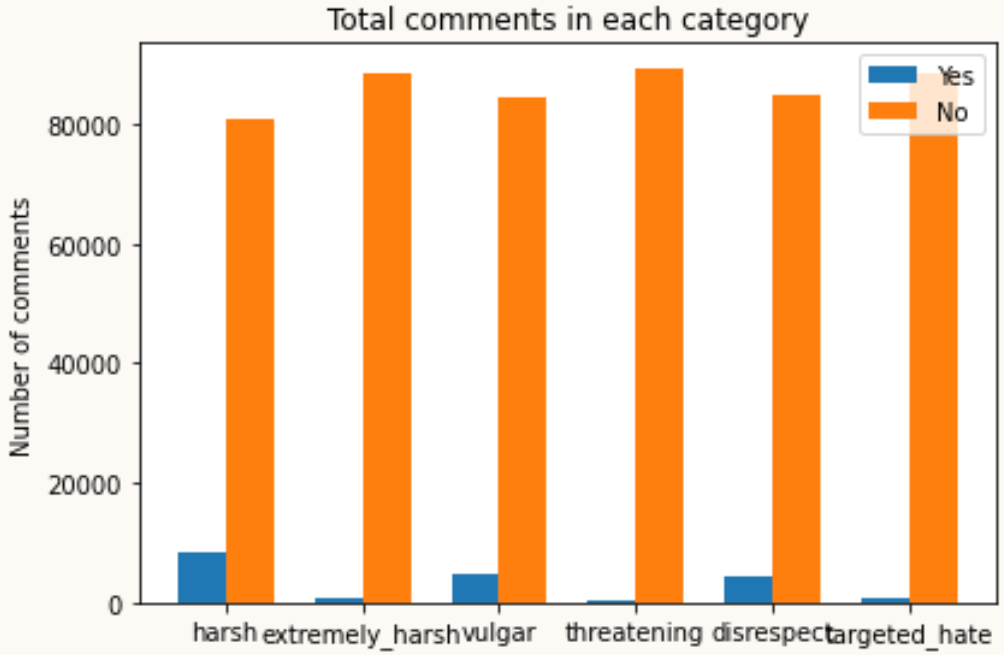
WHY SO HARSH

The given problem statement is to classify a multi-labelled dataset of comment texts sourced from discussion forums with the techniques of natural language processing (NLP).

EXPLORATORY DATA ANALYSIS

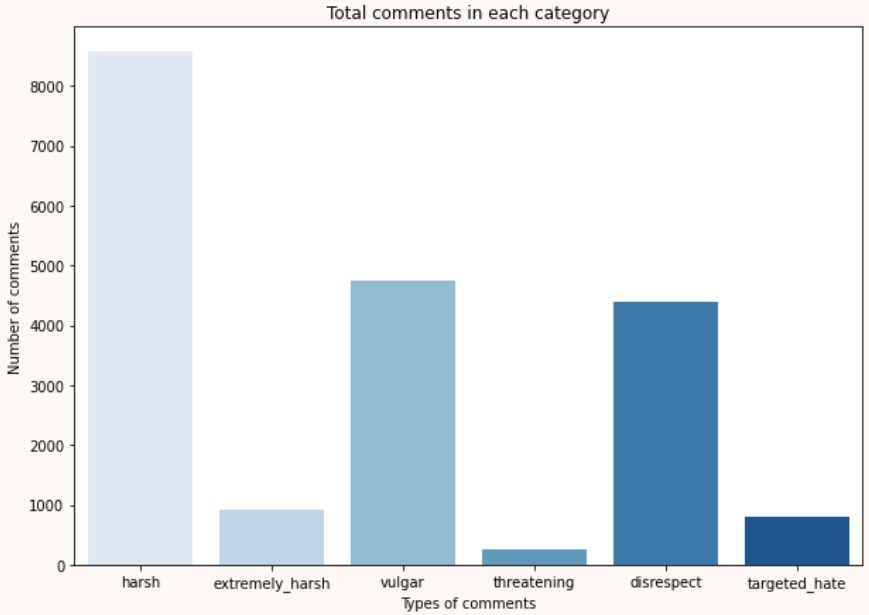
On exploring the comments given it was found to be highly skewed. The imbalance in the data comes from the fact that very few percentage of the comments are labelled with harsh, extremely harsh, etc.

As seen in the graph below an extremely low percentage of the comments are labelled as harsh or the other categories.



EXPLORATORY DATA ANALYSIS

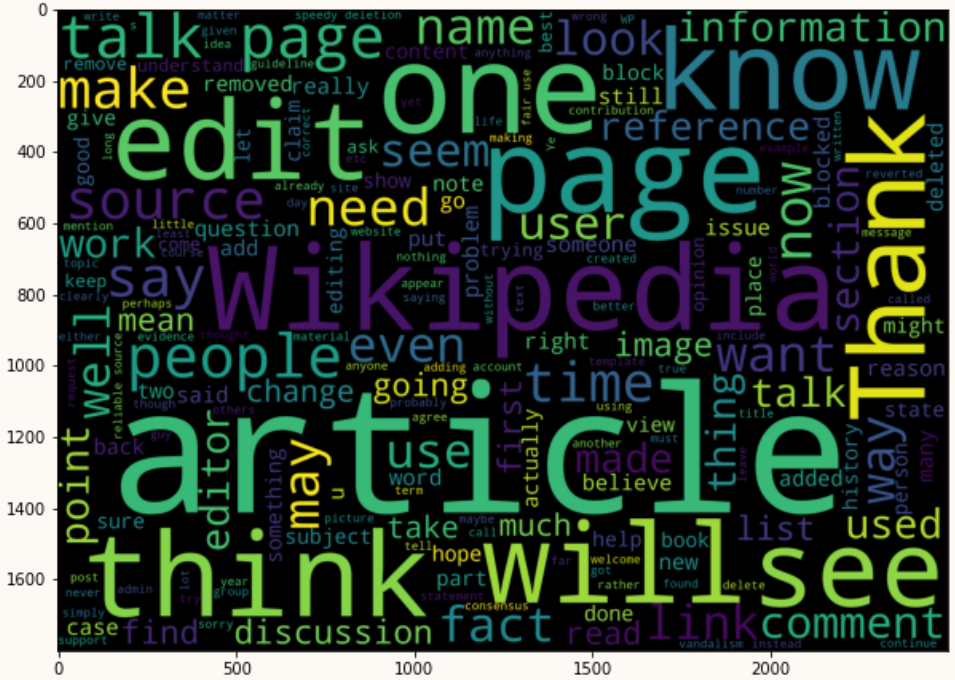
The count of total number of comments in each category is displayed in the graph below



As seen the number of comments is highest in the *harsh* category, while *threatening* category has the least number of comments.

EXPLORATORY DATA ANALYSIS

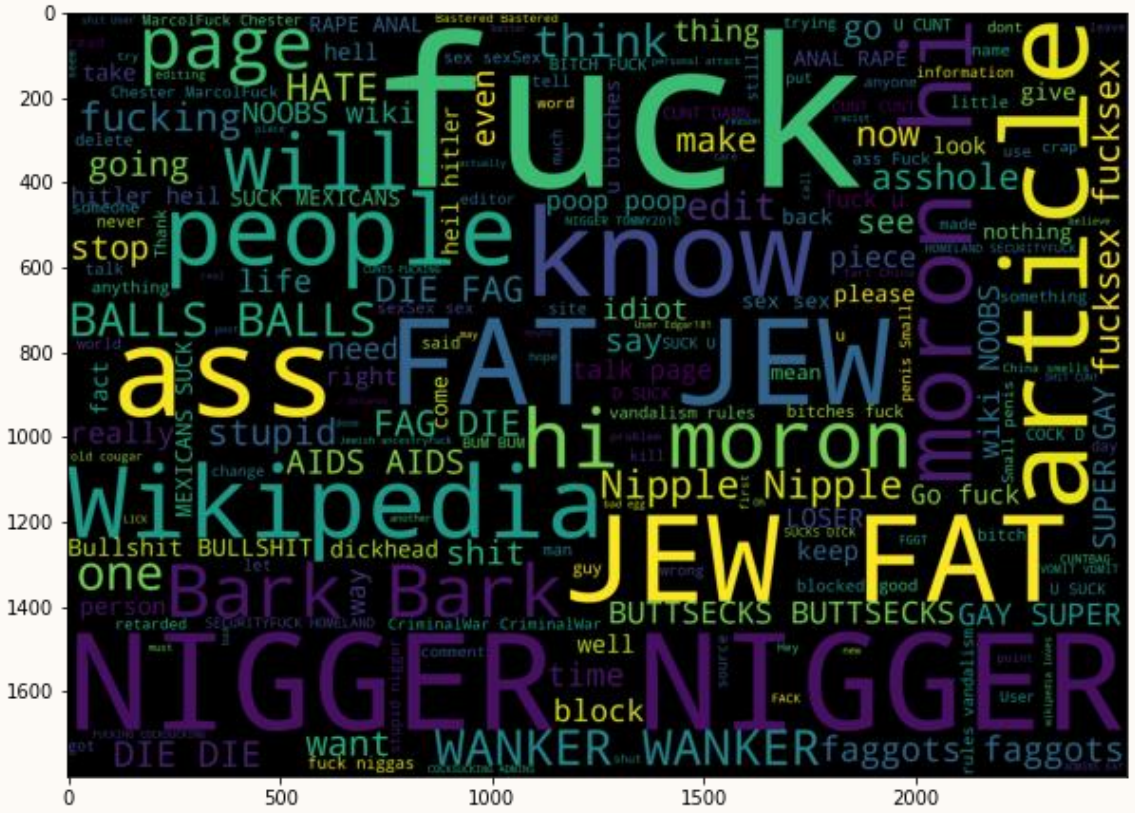
A WordCloud for the entire dataset was plotted and it is shown below.



Here, it is observed that most of the words in the corpus are regarding articles and discussions on the Wikipedia forum. This was used as a reference to select the glove word embeddings file used later.

EXPLORATORY DATA ANALYSIS

Next, a WordCloud only for the labelled comments was selected and plotted which looks as follows.



TEXT CLEANING

The following unnecessary text was removed from the given dataset of comments:

- URLs and HTML links
- Emojis
- Word contractions
- IP addresses

The letters were not converted to lowercase because in many of the comments, harshness is decided by the presence of Uppercase in the comment. In some cases, the presence of punctuation marks, such as the exclamation mark might also play a role in the classification of the comment.

LEMMATIZATION AND STOPWORD REMOVAL

Lemmatization was chosen for pre-processing of text rather than stemming, because it could retain the sentiment of a word, which would be more likely to be lost upon stemming of the word.

The standard set of English stop-words were selectively removed. Most of the punctuations were also removed during the process.

All the pre-processing was performed with the help of the NLTK library.

WORD VECTORIZATION USING TF-IDF VECTORIZER

TF-IDF vectorizer was chosen instead of Count vectorizer because it considers not only the frequency of words like the count vectorizer, but also provides certain importance to each word in the corpus.

Two types of vectorizers were combined to make the final vectorizer for the comment text –

1. Word vectorizer – Features are made of word n-grams
2. Character vectorizer – Features are made of character n-grams

This bit of vectorization improved the accuracy by a large percentage.

GLOVE VECTOR EMBEDDINGS

Pretrained word embeddings implemented by Stanford university called Global Vectors (GloVe) were used.
The glove.6B.zip file provides word vectors trained on data from Wikipedia 2014 forums and English Gigaword containing newswire text data.
This was thought to be the most relevant in our case as the word vectors were pretrained on Wikipedia data, from where our dataset was acquired.
The training data comments were then transformed and vectorized with these word embeddings. This was fed into a logistic regression model with hyperparameter tuning using the *sklearn* library *GridSearchCV*. But as it takes a very long time to train it is not feasible to be used for our purpose.
Cross validation accuracy score of the logistic regression model with GloVe vectors was obtained to be **93.3965%.**
Cross validation accuracy score of the Multinomial Naïve Bayes model with GloVe vectors was obtained to be **85.4036%.**

SAMPLING OF DATA

During EDA analysis of data, we observed that the data was very biased for all the labels. For some classes it was even below 1 percent of the entire data. Due to this, the model may never be able to predict against the bias. Thus, we tried **sampling** of the data to overcome this.

This method allows us to create duplicate samples of the minority classes. There are two types of sampling methods:

- **Over – Sampling** : By this method, we can decrease the bias by creating duplicate values of the minority classes for every label. This allows us to increase the bias and increase the size of training data.
- **Under – Sampling** : We can decrease the number of samples of the majority classes for the data. This would reduce the bias but would also reduce the size of the training data.

Oversampling using the SMOTE function was tried for the classes with worst bias. But this did not result in a better accuracy of our model hence we did not proceed forward with the balanced dataset.

FINAL MODELS IMPLEMENTED

The final models that were implemented using the TF-IDF pre-processed and vectorized data as input are as follows:

- Logistic Regression
- Multinomial Naïve Bayes

The cross validation score received for each of these models are listed in the table below:

Model Implemented	Logistic Regression	Multinomial Naïve Bayes
CV score (harsh)	97.6615	93.1769
CV score (extremely_harsh)	98.7621	91.8077
CV score (vulgar)	98.9274	92.7951
CV score (threatening)	98.9657	78.5636
CV score (disrespect)	98.1299	92.4060
CV score (targeted_hate)	98.0661	84.3459
CV score (Total)	98.4188	88.8492

SUMMARY

Finally, the models that were tested out were using different ore-processing techniques, specifically TF-IDF vectorization and GloVe word embeddings.

Out of these, the model which gave the best accuracy was the Logistic Regression model with TF-IDF vectorization with a cross validation score of 98.4188% and a Kaggle score of 98.541%.

REFERENCES

1. https://scikit-learn.org/stable/user_guide.html
2. https://nlp.stanford.edu/projects/glove/
3. https://github.com/lazyprogrammer/machine_learning_examples/blob/master/nlp_class2/bow_classifier.py
4. https://www.kaggle.com/code/thousandvoices/logistic-regression-with-words-and-char-n-grams/script
5. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
6. https://medium.com/thecyphy/handling-data-imbalance-in-multi-label-classification-mlsmote-531155416b87

THANK YOU

Aamod B K
Aamod.BK@iiitb.ac.in
Ishaan Jalan
Ishaan.Jalan@iiitb.ac.in