Laguna State Polytechnic University
Republic of the Philippines
Province of Laguna
Los Baños

Tweet Cyberbullying Classification

A Final Project Presented to the Faculty
of Computer Studies
of Laguna State Polytechnic University

In partial fulfillment of the requirements in
Natural Language Processing
Bachelor of Science in Computer Science-III A

By:
Gupo, Jhon Vincent A.

July 2022

INTRODUCTION

With the enormous amount of text data that we have today, text classification is widely used to sort the natural language data that we have today. To assess whether an incoming email is routed to the inbox or filtered into the spam bin, email software utilizes text categorization (Google Developers, 2021). Many companies also utilize this technique to evaluate customer feedback on their products or services to achieve a data-driven conclusion and reduce manual labor in manually reading reviews or feedback (Boukkouri, H., 2020).

Most people use social media as a vital form of communication daily, with usage rising across the board for all age groups. Due to social media's widespread use and relative anonymity, cyberbullying may negatively affect anybody at any time or location.
The internet makes it more challenging to resist such personal assaults than conventional bullying.

In reaction to the heightened risk of cyberbullying during the COVID-19 pandemic owing to extensive school cancellations, more screen usage, and less physical social interaction, UNICEF issued a warning on April 15th, 2020. 36.5 % of middle and high school kids have experienced cyberbullying, and 87 % have witnessed it. The effects of cyberbullying can include poor academic performance, despair, and suicide ideation. So, for our project presentation, we used a dataset from Kaggle for Cyberbullying Classification. It consists of approximately 47000 tweets classified in 5 labels which are the following:

- Religion
- Age
- Gender
- Ethnicity
- Not Cyberbullying

To check the dataset contents, we performed an Exploratory Data Analysis by checking some samples per category and performed a word cloud per category after preprocessing the tweets (such as stop words removal), which can be seen below.

TRIGGER WARNING! The following figures contain a lot of explicit words.

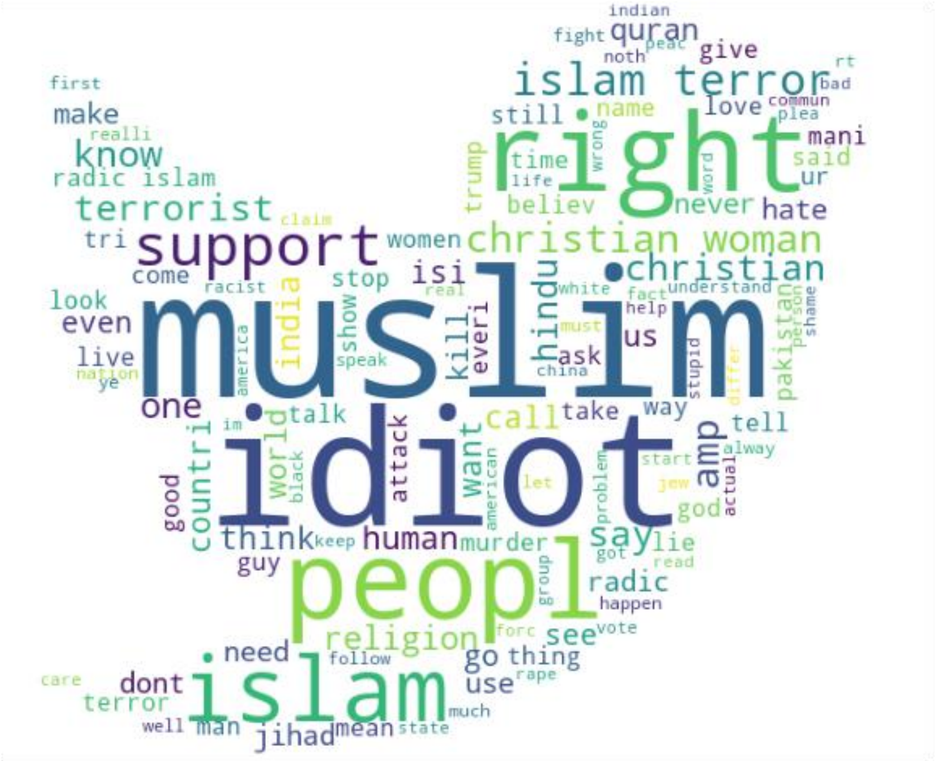

Figure 1. Religion Category

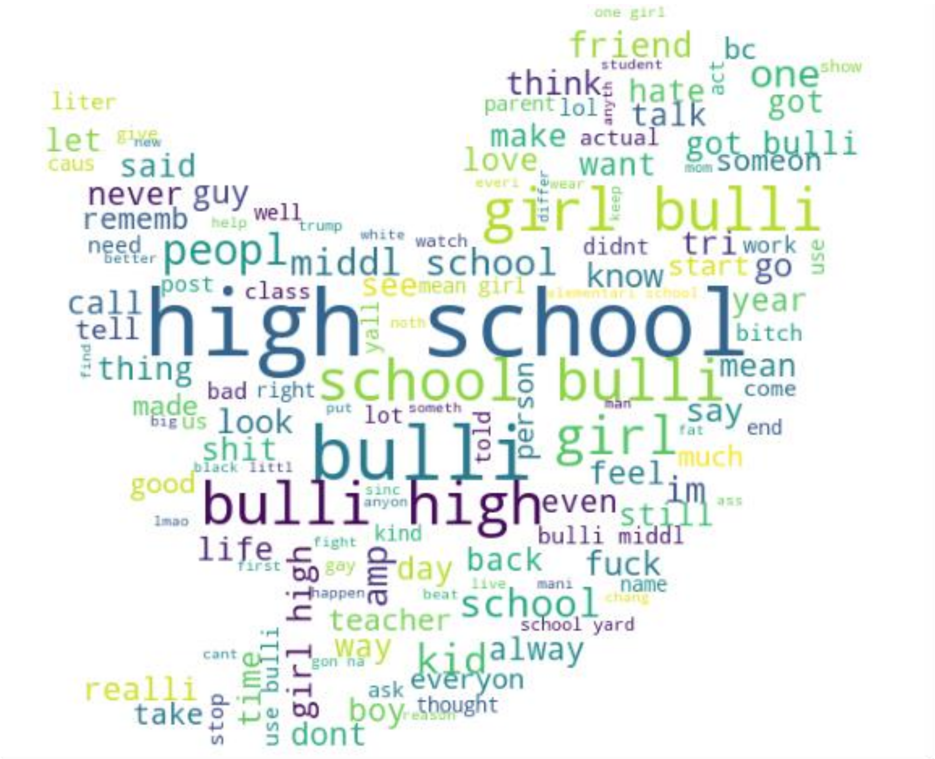

Figure 2. Age Category

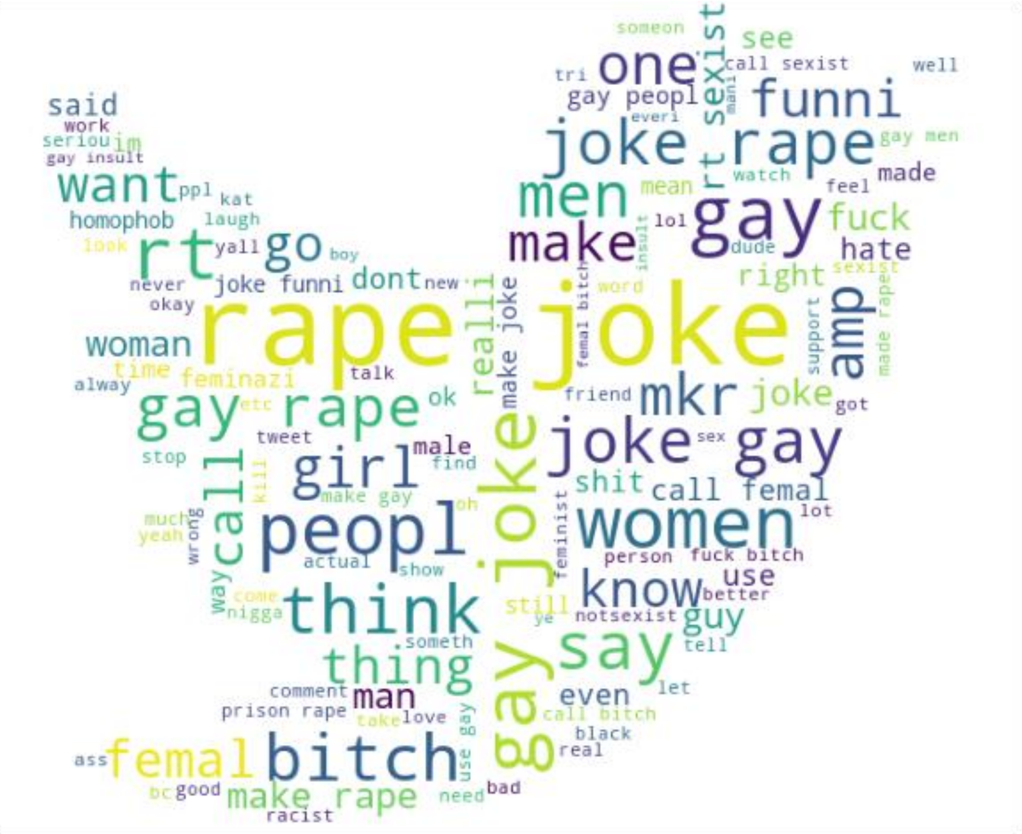

Figure 3. Gender Category

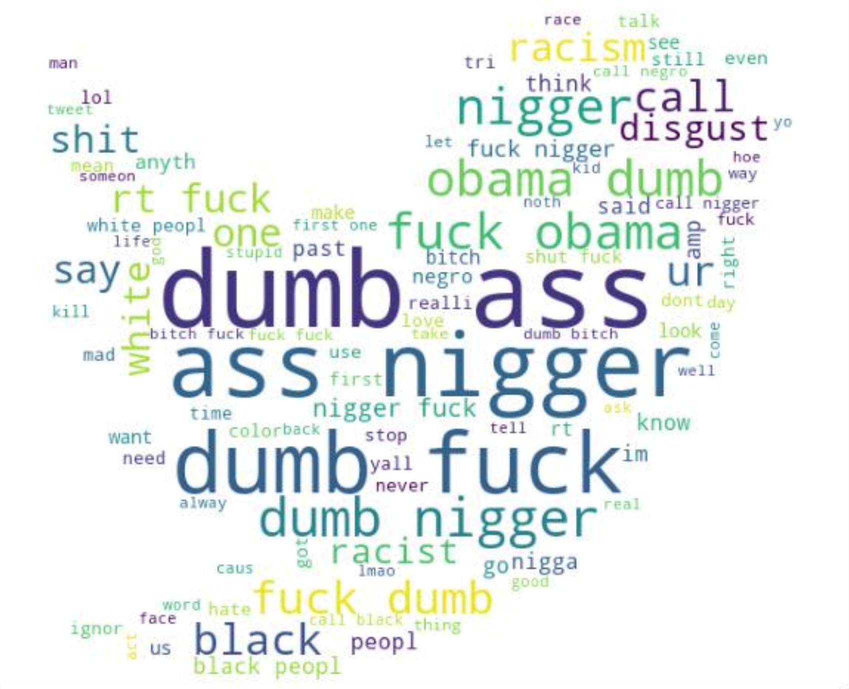

Figure 4. Ethnicity Category

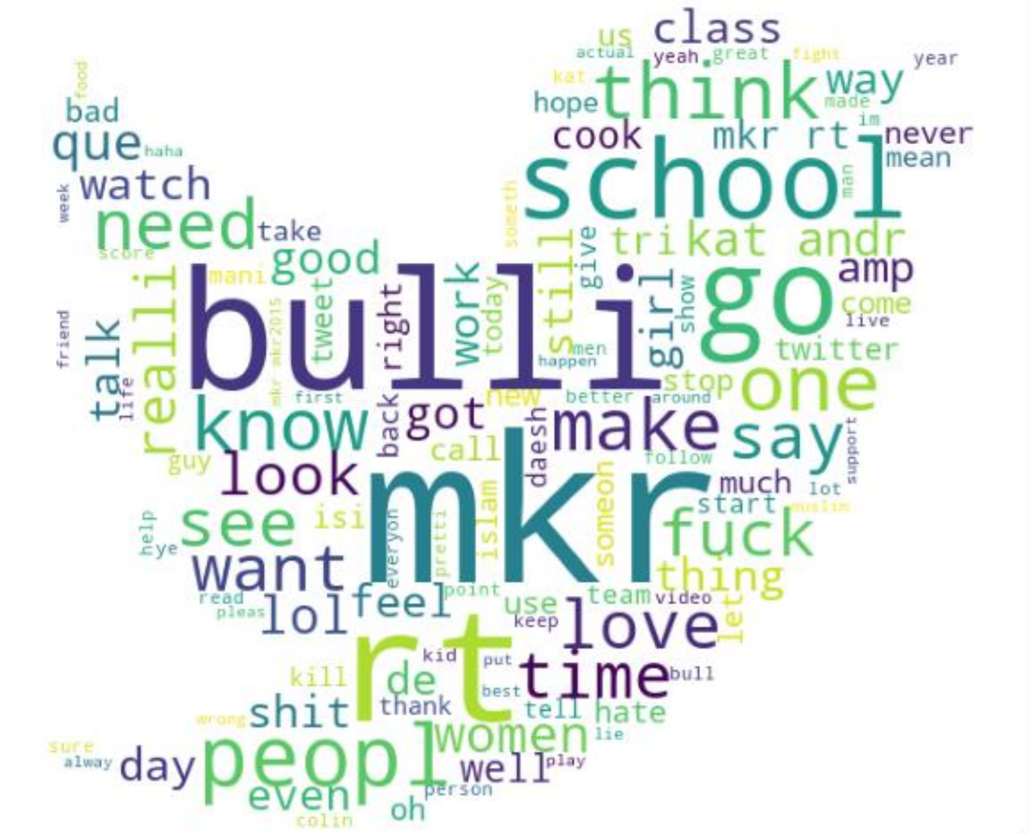

Figure 5. Not bullying Category

METHODOLOGY

The project pipeline was divided into five components: Data Processing, Feature Engineering, Dataset Splitting, Model Training, and Model Evaluation.

First, the data are preprocessed, such as removing stop words, removing emoji, normalizing letters for the features, and label encoding was done for the categories, and then the text is checked to remove duplication. The 2nd stage is the feature engineering to convert the tweets into numbers. Count Vectorizer was used to convert words to numbers, while TF-IDF was used for essential word extraction. The 3rd stage is splitting the data with 80:20 conversion for training and testing data with stratifying as a parameter for balanced class distribution. After checking the class distribution, there is approximately 1500 difference between the samples of the Religion and Not Cyberbullying category, so SMOTE is applied for dataset balancing. Three models were used to classify the categories which are Multinomial Naive Bayes, K Nearest Neighbor, and Random Forest. Lastly, the models are evaluated using the Classification Report for their F1 score, Precision, and Recall. Cross-Validation Score with three folds and visualizes each model's confusion matrix. For project pipeline visualization, you can refer at figure 6 at the next page.

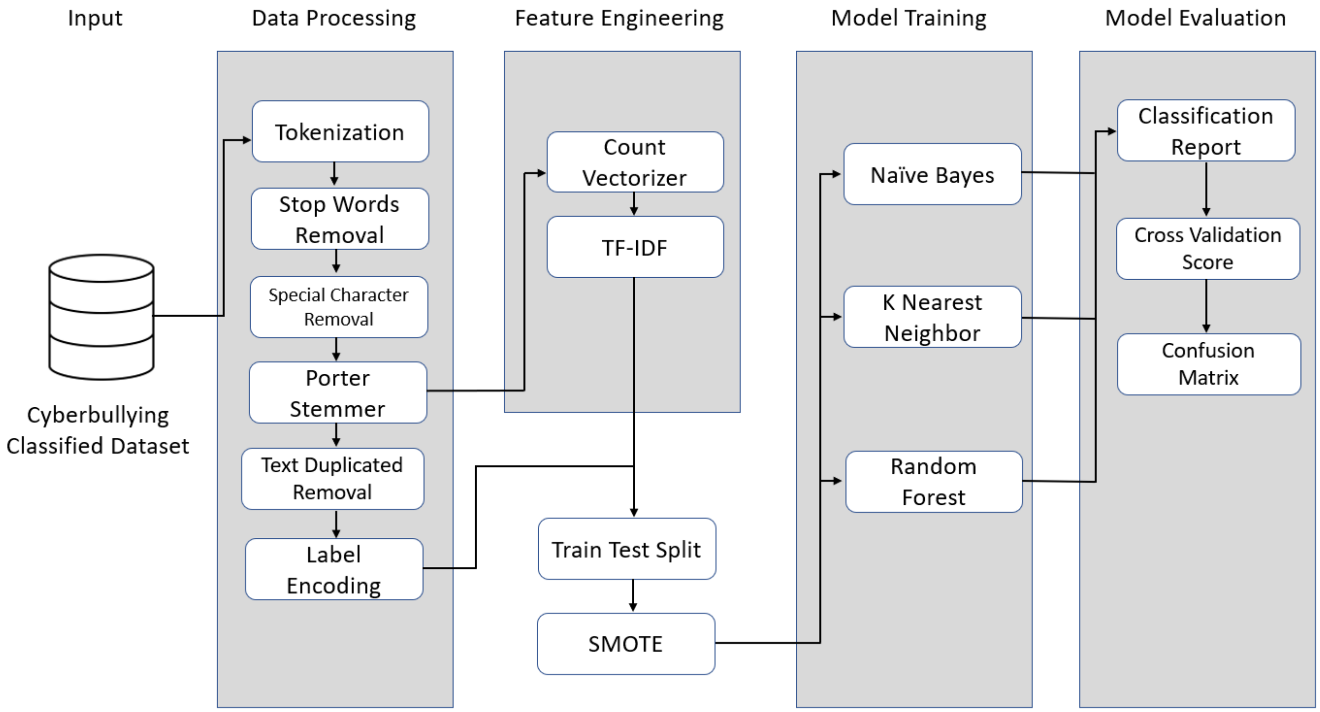

Figure 6. Project Pipeline

RESULTS AND DISCUSSION

Based on this study, Random Forests outperformed Naïve Bayes and KNN in all metrics, followed by Naïve Bayes in the 2nd place. Please refer to the notebook for each score per category. (Scores highlighted in red are the best scores)

Model	Precision	Recall	F1 Score
Naive Bayes	0.86	0.86	0.85
K Nearest Neighbor	0.81	0.69	0.70
Random Forest	**0.94**	**0.94**	**0.94**

Table 1. Classification Report Weighted Average

The result is also the same for cross validation score with Random Forest as the best classifier followed by Naïve Bayes.

Model	Score
Naive Bayes	0.849
K Nearest Neighbor	0.729
Random Forest	**0.939**

Table 2. Cross Validation Score

The next evaluation to be discussed is the confusion matrix. Here, we can see where categories each model misclassify at predicting.

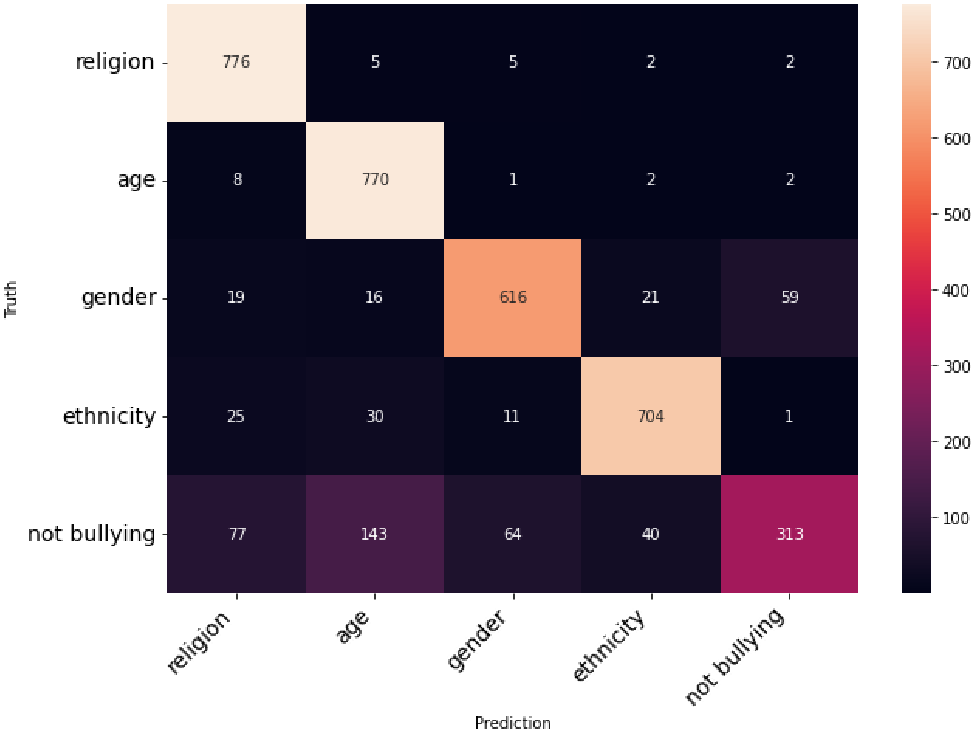

Figure 7. Naïve Bayes Confusion Matrix

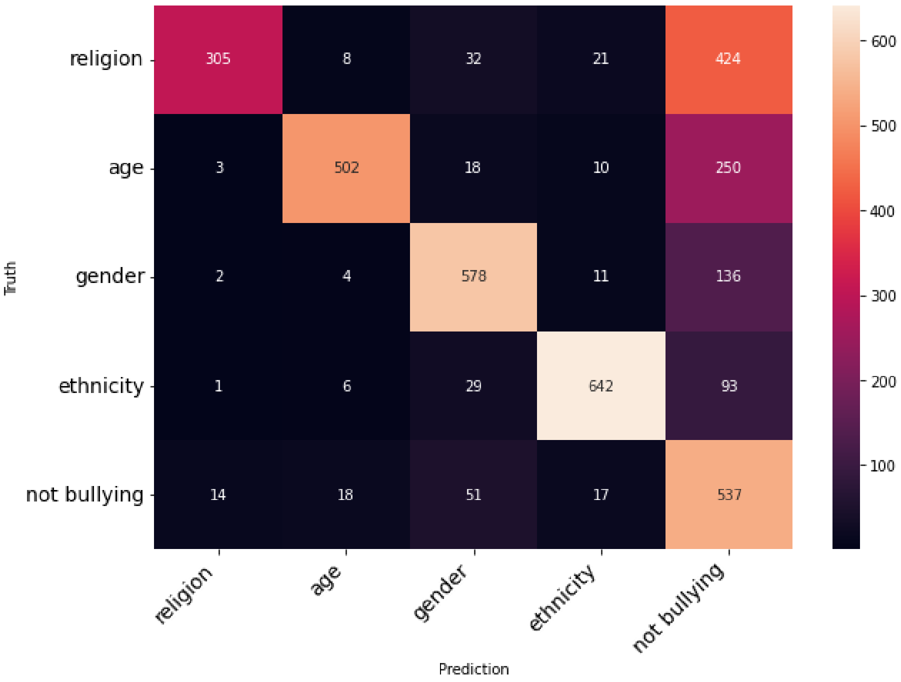

Figure 8. K Nearest Neighbor Confusion Matrix

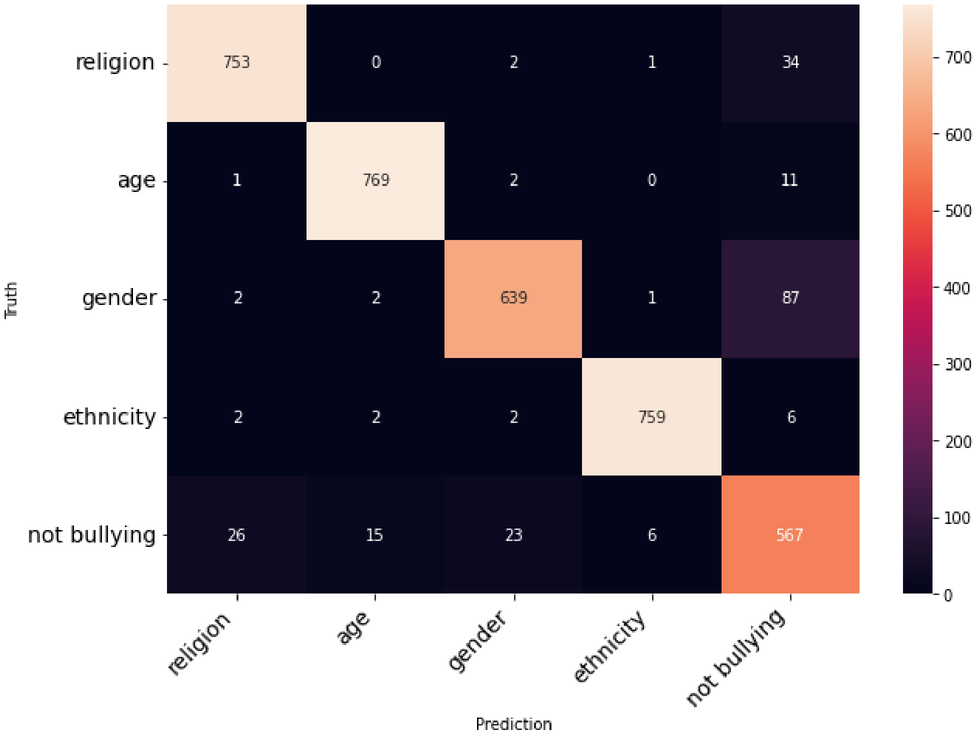

Figure 9. Random Forest Confusion Matrix

For Naïve Bayes, the not bullying category is the category that is the most misclassified while for KNN, the religion category is the most misclassified as not bullying and for Random Forest, gender category is the most misclassified category.

The next thing that we did was classify another dataset using the best classifier, which is Random Forest. For this dataset, we collected the data from 4 different datasets, which can be found in the notebook provided. The dataset was then preprocessed with the same pipeline and predicted using the best model.

Classification Report (weighted average)			
Random Forest	Precision	Recall	F1 Score
	0.74	0.66	0.61

Table 3. Another Dataset Classification Reports

Random Forest scores drop significantly with the test data. It seems like the words/ structure of the new data is not yet learned by the model

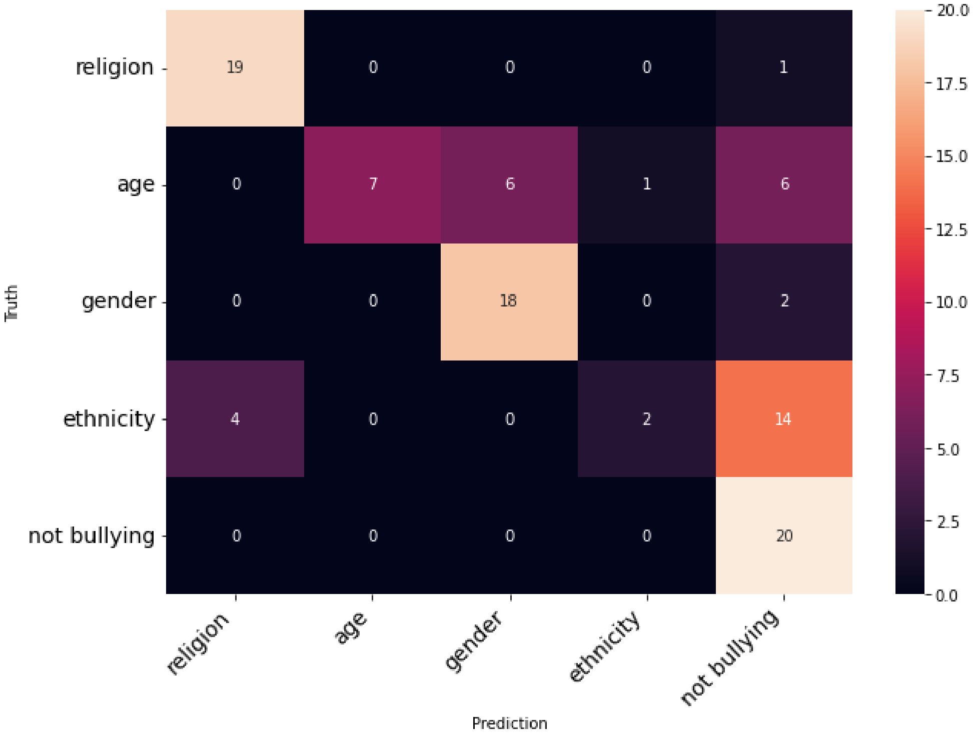

Figure 10. Test Data Random Forest Confusion Matrix

There are a lot of categories that are misclassified as not bullying. Perhaps there are words within the test data that are in the not bullying category of the training data.

Here are some of the results on the test data with respect to their original sentiments and random forest predicted sentiments.

Text	Sentiments	Predicted
I probably would not mind school as much if we did not have to deal with bitch ass teachers". Retweet	Age	Age
CauseWereGuys: On my way to fuck yo bitch me as a 9 year old	Age	Age
The PKI only appears when the regional head and presidential elections are approaching because they are fried until they are burnt.	Ethnicity	Not Cyberbullying
RT @user I'm not sexist but some things shouldn't be said by females	Gender	Gender
Kat I'd love to slap your face with a pork cutlet	Gender	Gender

Islam is not a race, microbrain, it is a death cult.	Religion	Religion
i still can not wrap my head around the fact that #christinagrimmie is gone. and the fact that a man destroyed #prayfororlando just	Not Cyberbullying	Religion
i am thankful for cats. #thankful #positive	Not Cyberbullying	Not Cyberbullying

Table 4. Predicted Sentiment of Random Forest on new dataset

If you want to explore more text with respect to their sentiments and predicted sentiments from Random Forest, please refer to the Jupyter Notebook included in the attachment.

CONCLUSION

Based on this study, Random Forest performed the most accurately in both evaluation metrics, followed by Multinomial Naïve Bayes and K Nearest Neighbor at the lowest. However, the Random Forest score drops significantly in testing a dataset from another source. It might be because of new words appearing in the other dataset that the model did not learn. There are also a lot of misclassified categories that Random Forest predicts as not bullying. Further deep-dive analysis of the not bullying category in the train dataset and retraining some parts of the test dataset before the model predicts is recommended.

RECOMMENDATION

It is recommended to have the same research be done in the future that tweaks the following:

For the dataset, Future researchers can make other datasets with the same category as this to further expand the number of data to be trained and deep dive analysis of the not bullying category is recommended.

For data processing, another pipeline configuration can be used, such as using a Lemmatizer instead of a Stemmer. More preprocessing layers can be done to avoid words that don't have a meaning being fed into the models.

For feature engineering, other techniques such as Word2Vec or Word embeddings can be learned more here, and powerful language models such as Google's BERT, which is already trained on millions of texts and has contextual embeddings, that can be learned more here.

For Machine Learning models, Hyperparameter Tuning can be performed using Grid Search and you can also use deep learning models that specialize for natural language data such as LSTM or GRU.

Reference

Google Developers (2021, May 26). Text Classification Introduction available at https://developers.google.com/machine-learning/guides/text-classification

Boukkouri, H. (2020, August 25). Text Classification: The First Step Toward NLP Mastery available at https://blog.dataiku.com/text-classification-the-first-step-toward-nlp-mastery

Larxel (2022, January) Cyberbullying Classification dataset available at https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification