CYBERBULLYING DETECTION USING SENTIMENT ANALYSIS

Nethmi Fonseka

I. INTRODUCTION

The In the era of high social media usage, the prevalence of cyberbullying has emerged as a critical societal issue. The omnipresence of social media platforms implies that cyberbullying can impact individuals of all ages, transcending geographical and temporal boundaries. UNICEF's warning on April 15th, 2020, highlighted the escalating risk of cyberbullying during the COVID-19 pandemic. Widespread school closures, increased screen time, and reduced face-to-face social interaction have created a fertile ground for digital harassment. The statistics reveals 36.5% of middle and high school students experiencing cyberbullying and its impacting academic performance, mental health, and even leading to suicidal thoughts.

The project proposes an approach to identify cyberbullying through advanced data analysis using natural language processing and machine learning techniques. I have used a labeled dataset from Kaggle which includes over 47,000 labeled tweets, provides a rich source of information on cyberbullying across dimensions such as age, ethnicity, gender, religion, and different types of cyberbullying.

By categorizing instances based on age, ethnicity, gender, religion, and specific types of cyberbullying, aims to provide targeted insights for effective intervention strategies. Additionally Employing sentiment analysis and behavioral pattern recognition, this system will serve as a proactive tool, flagging content indicative of cyberbullying and enabling timely and targeted responses.

II. METHODOLOGY

This chapter outlines the methodology employed, encompassing data collection, data processing, preliminary data analysis, and the application of relevant statistical using Natural Language Processing and Machine Learning techniques.

Data Collection

The dataset utilized in this study was obtained from J. Wang, K. Fu, C.T. Lu's work titled "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection" and is publicly available on Kaggle at the following link: [Cyberbullying Classification Dataset] (https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data).

```
df = pd.read_csv('/content/cyberbullying.csv', nrows = 20000)
df
```

	tweet_text	cyberbullying_type
0	I would back hand them both. Giving them a 1? ...	not_cyberbullying
1	Niggas wanna fight after school and shit	not_cyberbullying
2	#mkr say sassy one more time 😊	not_cyberbullying
3	Can't believe I'm still awake when I have to ...	not_cyberbullying
4	@OdiniaInvictus @SirajZarook @BilalIGhumman @I...	not_cyberbullying
...	...	...
19995	you ignorant BITCH, Suck & Black Dick Your Whi...	ethnicity
19996	and your body goes oh shit nigger what the fuc...	ethnicity
19997	@FsN_Vigsta fuck that dumb nigger	ethnicity
19998	...did a white woman really just send me a vid...	ethnicity
19999	I fully agree But lately, it's become justifie...	ethnicity

20000 rows × 2 columns

Figure1 Dataset

Due to lack of GPU in Google Colab, a subset of the original dataset containing 20,000 rows was selected for analysis. The dataset includes information about tweets and their corresponding labels, categorized based on the class of cyberbullying, including age, ethnicity, gender, religion, other forms of cyberbullying, and instances not classified as cyberbullying.

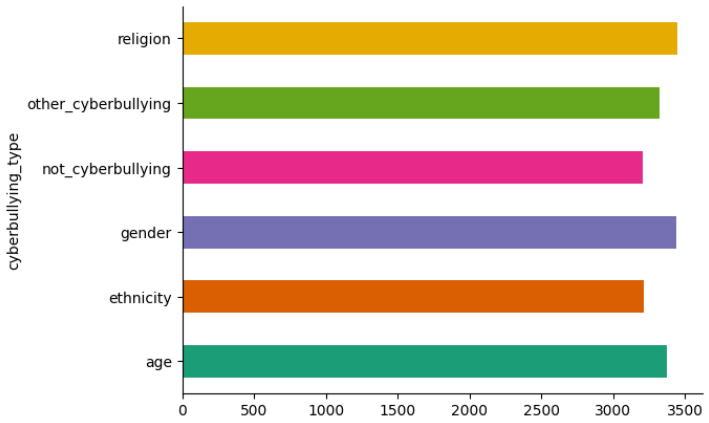

Figure2 Categories of dataset

Data Processing

Prior to analysis, data cleaning procedures were implemented to ensure the quality and relevance of the information. This involved handling missing values, checking for duplicates, and cleaning the text data by removing HTML tags, non-alphabetic characters, URLs, and user mentions. Additionally, the text was tokenized, stopwords were removed, and words were stemmed to enhance the effectiveness of subsequent NLP techniques.

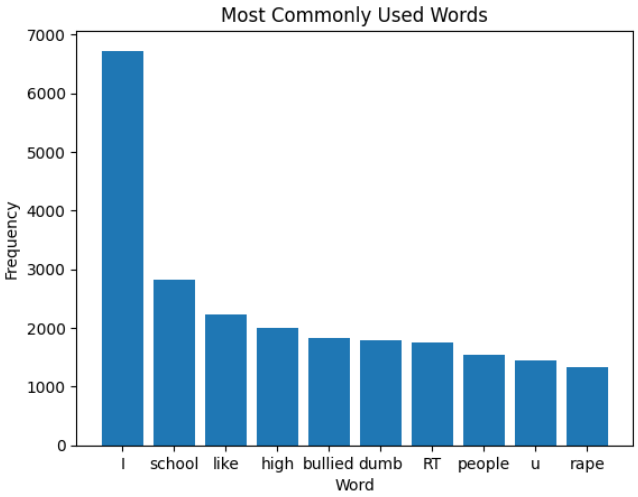

Figure3 Frequently Used Words

Preliminary Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain insights into the distribution and characteristics of the dataset. This encompassed visualizations of the class distribution of cyberbullying types, sentiment analysis plots, and frequency analysis of the most commonly used words. EDA allowed for a deeper understanding of the dataset's structure and potential patterns.

To extract relevant statistics, the text data underwent further processing to remove stopwords and non-alphabetic characters. The frequency of each word was then analysed, leading to the identification of the most commonly used words in the dataset. Sentiment analysis was performed using the TextBlob library to generate polarity and subjectivity scores, providing additional insights into the emotional tone of the tweets.

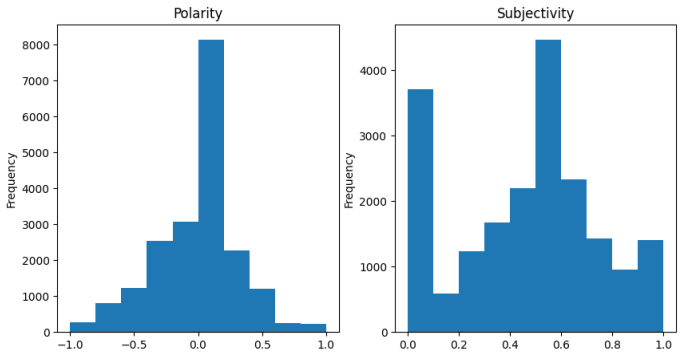

Figure 4 Sentiment Analysis Plot

The histogram of polarity scores shows the distribution of sentiment polarity in the dataset. The sentiment polarity ranges from -1 (indicating a highly negative sentiment) to 1 (indicating a highly positive sentiment), with 0 representing a neutral sentiment. The histogram visually demonstrates how many tweets fall into different sentiment polarity categories.

The histogram of subjectivity scores illustrates the distribution of how subjective or objective the tweets are. Subjectivity values range from 0 (indicating a highly objective statement) to 1 (indicating a highly subjective or opinionated statement). The histogram visually conveys the spread of subjectivity levels across the dataset.

For both plots, higher bars indicate a higher concentration of tweets with specific sentiment or subjectivity scores.

The distribution of polarity scores provides insights into whether the tweets lean towards positive, negative, or neutral sentiments.

The subjectivity histogram helps understand the degree of objectivity or subjectivity present in the dataset.

III. RESULTS

I chose Logistic Regression model to train the dataset. I got an Accuracy Score of 0. 82325 by training this dataset.

The provided confusion matrix in the Google Colab represents the performance of a

classification model on a multiclass problem with six classes. Each row of the matrix corresponds to the true class, while each column corresponds to the predicted class. The numbers in the matrix indicate the counts of instances for each combination of true and predicted classes.

The diagonal elements (from the top-left to bottom-right) represent correct predictions. Off-diagonal elements represent misclassifications. The values in each cell indicate the count of instances for the corresponding combination of true and predicted classes.

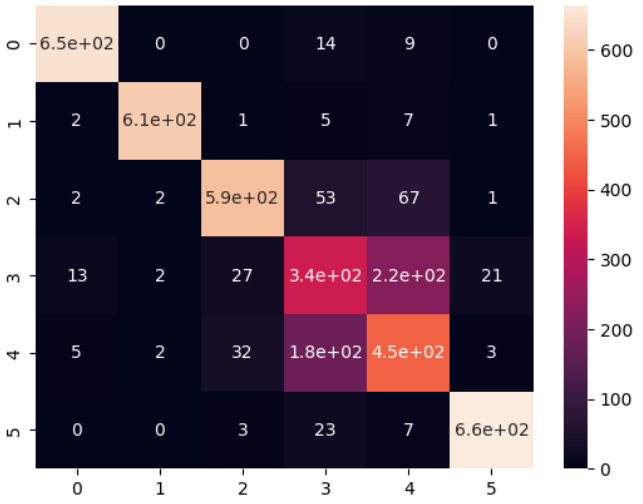

Figure5 Heatmap for Confusion Matrix

For example, the element in the first row and first column (index 0, 0) with a value of 648 indicates that there were 648 instances where the true class was 0 (Class 0), and the model correctly predicted it as Class 0.

The element in the fourth row and fifth column (index 3, 4) with a value of 221 indicates that there were 221 instances where the true class was 3 (Class 3), but the model incorrectly predicted it as Class 4.

So, the general observation of the Confusion Matrix is that the model generally performs well, as evidenced by the high values on the diagonal.

```
from sklearn.metrics import classification_report
report = classification_report(y_test, y_pred)
print(report)
```

```
                    precision    recall  f1-score   support

               age       0.97      0.97      0.97       671
         ethnicity       0.99      0.97      0.98       628
            gender       0.90      0.82      0.86       713
not_cyberbullying       0.55      0.54      0.54       619
other_cyberbullying       0.59      0.66      0.62       672
          religion       0.96      0.95      0.96       697

          accuracy                           0.82      4000
         macro avg       0.83      0.82      0.82      4000
      weighted avg       0.83      0.82      0.83      4000
```

Figure6 Classification Report

The model achieves an overall accuracy of 82%, indicating that it correctly predicts the class labels for approximately 82% of the instances in the test set.

Class-Specific Performance

The precision, recall, and F1-score for each class provide insights into how well the model performs for individual classes.
For instance, the model demonstrates high precision and recall for the "age" and "ethnicity" classes, indicating robust performance.
The "not_cyberbullying" class has lower precision and recall, suggesting challenges in correctly identifying this class.

Macro and Weighted Averages

Macro average calculates the unweighted mean of precision, recall, and F1-score across all classes. It provides an equal contribution to each class.
Weighted average considers the number of instances in each class and provides a weighted mean, giving more importance to larger classes.
Overall Assessment:

The model exhibits good performance on certain classes but faces challenges in distinguishing the "not_cyberbullying" and "other_cyberbullying" classes.The weighted average F1-score of 0.83 indicates a reasonable balance between precision and recall across all classes, considering class imbalances.

CONCLUSION

Developing a cyberbullying detection system using Natural Language Processing (NLP), the project has yielded valuable insights into the dynamics of cyberbullying within a social media dataset. The analysis involved comprehensive data processing, exploratory data analysis (EDA), sentiment analysis, and the implementation of a classification model.

The Logistic Regression model exhibited commendable performance with an overall accuracy of 82%. This indicates a reliable ability to classify tweets into different categories of cyberbullying. The confusion matrix and classification report provided a detailed understanding of the model's strengths and weaknesses across various classes. High precision and recall were achieved for certain classes, while others presented challenges.

The project's structured methodology and the availability of well-documented libraries facilitated the implementation process. Leveraging tools like scikit-learn and TextBlob streamlined key tasks, from data processing to model training.

Handling class imbalances and achieving a balanced performance across all classes proved challenging. The model exhibited varying degrees of success in distinguishing certain types of cyberbullying, highlighting the inherent complexities of identifying nuanced patterns in text data.

Addressing the challenges observed in the model's performance, particularly in classes with lower precision and recall, requires further investigation. Fine-tuning hyperparameters, experimenting with alternative algorithms, and incorporating additional relevant features could lead to improved results.

Exploring more advanced NLP techniques, such as deep learning models, may uncover more intricate patterns in cyberbullying language. Additionally, expanding the dataset and refining the model with continuous learning could enhance its adaptability to evolving online communication trends.

REFERENCES

J. Wang, K. Fu, C.T. Lu, “SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection,” Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), December 10-13, 2020.