Report on

CA2 - Text Mining

By

Kashmira **20023193**

Submitted to

Lecturer **Kunwar Madan**

Report on Text Processing and Model Evaluation

Introduction

This report critically analyzes the steps involved in text cleaning, the creation of structured data from text, and the evaluation of model performance based on a provided code implementation. The dataset used is a collection of tweets, with a focus on detecting cyberbullying. The key steps include text preprocessing, transforming text data into a numerical format, and evaluating two machine learning models: Support Vector Classifier (SVC) and Naive Bayes Classifier (NBC).

Text Cleaning

The text cleaning process involves several steps to preprocess the raw tweet data before converting it to a numerical format and following steps are performed for the process

1. **HTML Tag Removal**: BeautifulSoup is used to eliminate HTMLtags that are present in the tweets, to ensure that the text is free from web-specific tags that can cause issues while execution.
2. **Removing Mentions and URLs:** Using regular expressions to substitute mentions (@), URLs (http://, https://, www), and unwanted characters (e.g., \\x) with whitespace.
3. Non-alphabetic Character Removal: Have substituted any non-alphabetic characters with whitespace using regular expressions.
4. **Tokenization**: Tokenized the cleaned text using NLTK's word_tokenize function and have split sentences into constituent words.
5. **Lowercasing**: Have Converted all tokens to lowercase.
6. **Top Words Removal**: Filtered out common English stop words using NLTK's stopwords list.
7. **Lemmatization**: Converted plural forms to singular forms

Limitations

- Information Loss: Removing non-alphabetic characters and stop words can result in the loss of potentially relevant information, such as emphasis or sentiment expressed by punctuation or particular stop words.
- Regular Expressions Complexity: The correct application of regular expressions may sometimes result in overfitting particular patterns, missing unwanted patterns, or in error deleting text.

Creation of Structured Data

After cleaning the text, the next step performed, converted the text into a structured numerical format using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. The steps involved are:

1. TF-IDF Vectorization: Transforming the cleaned text data into TF-IDF features using scikit-learn's TfidfVectorizer.
2. Vocabulary Export: Saving the TF-IDF vocabulary to a CSV file for future reference.

Limitations

- Sparsity: TF-IDF can generate a sparse matrix, particularly with a big vocabulary, which can be technically and physically costly.
- Context Ignorance: TF-IDF does not capture word context or order, therefore semantic details in the text may be missed.
- Fixed Vocabulary: Once developed, the vocabulary becomes fixed, making it impossible to deal with new words or terms that may occur in future data.

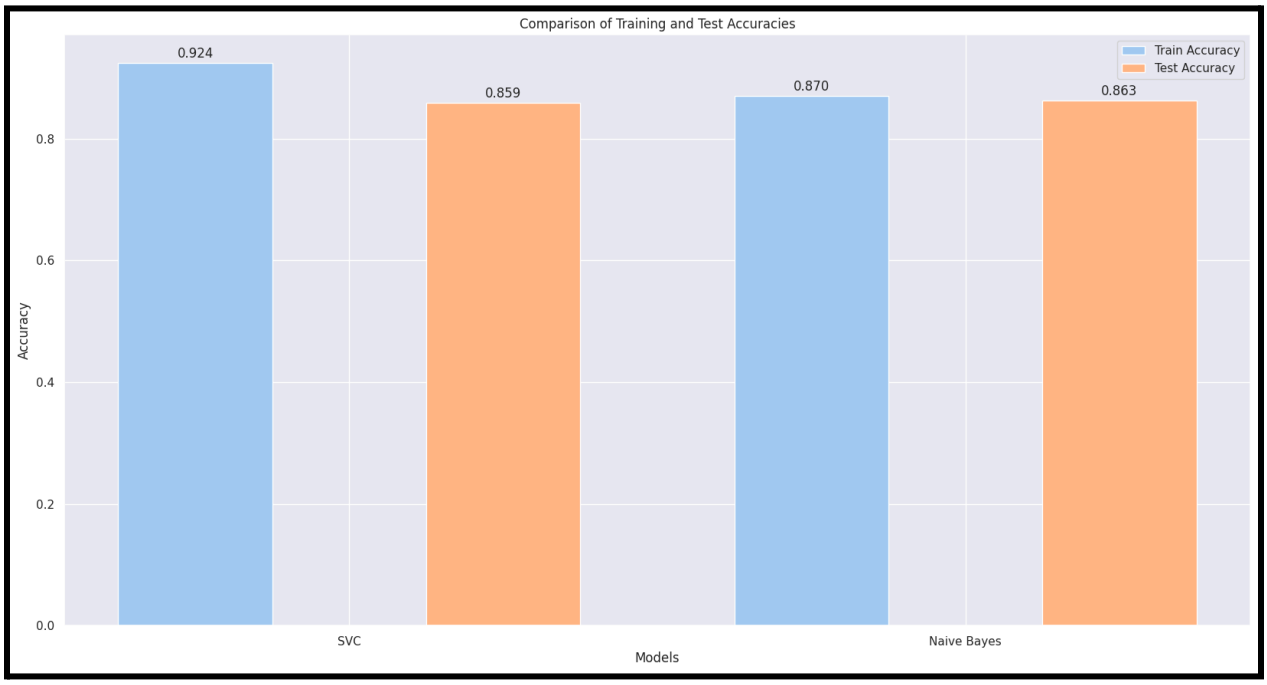
EDA - Exploratory Data Analysis

I have also created some visualizations to understand and explore data in an analytical and creative way.

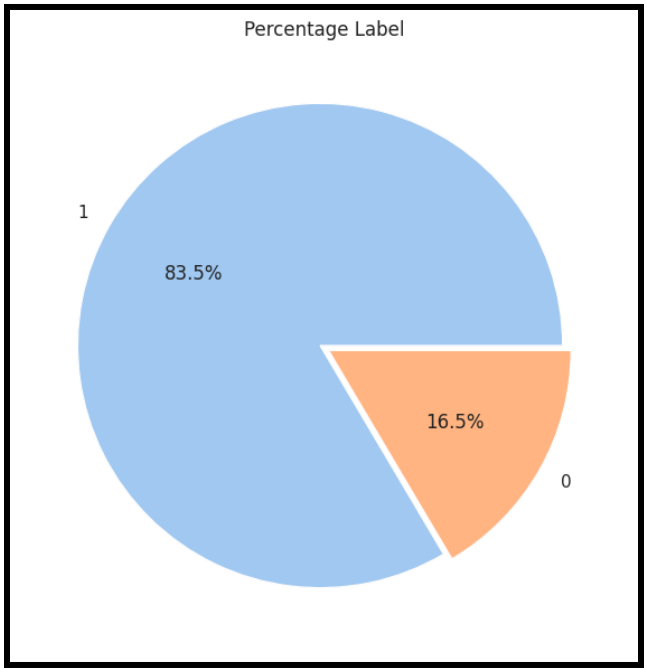
Following Class Distribution is done to create a visualization

- The cyberbullying_type column is mapped to binary values, with various types of cyberbullying assigned a value of 1 and non-cyberbullying assigned a value of 0.
- The class distribution is examined by counting the occurrences of each label (cyberbullying and not_cyberbullying).

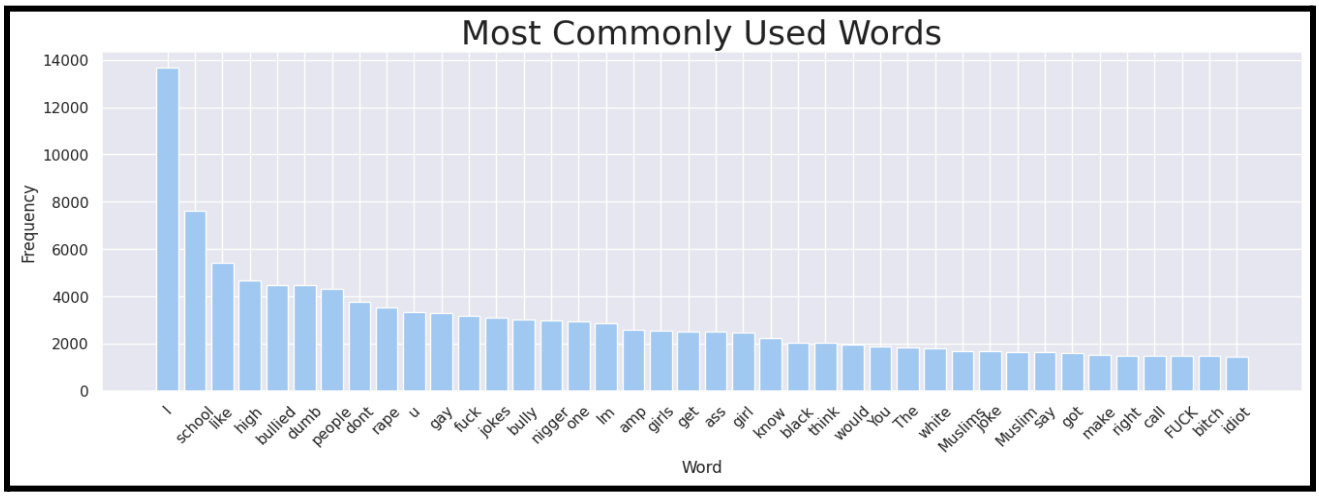
- Bar chart - It is showing the comparison between SVC and Naive Bayes displaying the train and test accuracy column chart for both models



- A pie chart is created to visualize the percentage distribution of each label in the dataset.



- Below chart displays the graph of most commonly used words in the tweets and the frequency of times that tweet is used.



Model Performance Evaluation

Evaluated two models using 10-fold cross-validation: Support Vector Classifier (SVC) and Naive Bayes Classifier (NBC).

1. Support Vector Classifier (SVC):
    a. Used LinearSVC with a linear kernel and default regularization parameter (C=1).
    b. Evaluated using 10-fold cross-validation, computing accuracy for each fold, and calculating the mean precision
2. Naive Bayes Classifier (NBC):
    a. Used MultinomialNB as it most commonly used for text classification also it gives a better result for text classification
    b. Similar evaluation procedure is used as SVC, 10-fold cross-validation and computed the mean precision.

Evaluation Metric: Precision

Precision was chosen as the key evaluation criteria, rather than accuracy as the dataset given was imbalanced so precision would give better results than accuracy. Precision is the ratio of true positive predictions to all positive predictions made by the model. This is particularly significant in the context of cyberbullying detection for the following reasons:

- False Positives Minimization: In cyberbullying detection, a false positive (incorrectly identifying a non-cyberbullying tweet as cyberbullying) can have serious social and ethical consequences. Precision aims to reduce these false positives.
- Importance of the Correct Detection: Precision ensures that the model's positive predictions remain highly relevant, which is critical for useful information in sensitive situations such as cyberbullying detection.

Model Performance Interpretation

Both SVC and Naive Bayes showed approx similar precision rate a slight difference of .4 was observed

- Mean cross-validation precision: **0.775**
- Naive Bayes - Mean cross-validation precision: **0.779**

Recommendations

Both models performed similarly, making them suitable for real-world use. SVC may be better for stable performance across varied datasets, as it can handle high-dimensional data more successfully. Tuning hyperparameters, using advanced text representation techniques such as Word Embeddings or BERT, and considering ensemble methods can all lead to improved performance. When compared the train and test accuracy, the test accuracy for Naive Bayes showed a value of 0.863 which is higher than SVC (0.859)
Considering the greater test accuracy Naive bayes is a better suggested model than SVC for text classification also its precision value is slightly higher than SVC

Conclusion

To conclude, I have mentioned critical analysis of the text cleaning process, the creation of structured data, and model performance evaluation. The approaches used, their limitations, and the performance of the SVC and NBC models are being explained. Both models demonstrated similar precision values ,Naive Bayes being slightly higher preferred for probabilistic models commonly used for text classification.Hence showing good performance in test accuracy and precision score, for the detection of cyberbullying dataset best suggested model is Multinomial Naive Bayes Classification