CIS 600: Social Media and Data Mining
Project Report

"Multi-Model Social Media Hate Speech and Cyberbullying Detection"

Submitted By:

1. Aatman Patel (SUID: 523227053)
2. Darsh Shah (SUID: 547308192)
3. Hemil Shah (SUID: 610010477)
4. Pavan Pandya (SUID: 340197894)
5. Samruddha Deshmukh (SUID: 379991386)

Table of Contents

1. Introduction

“Social media platforms have become widespread in recent years, which has brought about a number of advantages including unprecedented connectivity and information sharing." But there are some disadvantages to this technical breakthrough as well. A worrying trend that is overshadowing the benefits of internet connection is hate speech and cyberbullying. Since hostile rhetoric and targeted attacks can create an environment of animosity, these problems have an adverse effect not just on individual users but also on entire communities. It is necessary to address the negative parts of social media and investigate creative ways to lessen the potential harm that individuals and society may suffer if social media especially Twitter remains a prominent forum for public discourse.”

Hate speech is growing more common as information on the internet expands. We draw attention to and examine the difficulties in detecting hate speech in texts using online computer algorithms. Among the difficulties are linguistic nuances, varying perspectives on what hate speech is, and limitations on the amount of data that can be used to train and assess these algorithms. Moreover, a lot of contemporary methods suffer from understanding problems, which makes it challenging to comprehend why the systems behave the way they do.

“To address these issues head-on, this research develops and evaluates an advanced multi-model method specifically designed for the identification of hate speech and cyberbullying on Twitter and YouTube. The intricacy of social media communication necessitates a sophisticated approach that blends multiple models to improve precision and efficiency. The goal is to not only detect instances of hazardous information but also to understand the contextual nuances around such occurrences by exploring the complexities of language, sentiment, and user interactions. By taking this all-encompassing strategy, we hope to help build more inclusive and safe online environments, promoting better online interactions for those navigating the ever-changing social media landscape.”

2. Problem Statement and Approach

a. Problem Statement

- The advent of social media platforms has catalyzed a paradigm shift in communication, enabling global connectivity and diverse interactions. However, this surge in online engagement has brought to light a troubling facet— the proliferation of hate speech and cyberbullying. These issues pose significant threats to individual well-being, societal harmony, and digital community safety.
- Hate speech encompasses expressions that demean, threaten, or incite violence against individuals or groups based on their race, ethnicity, religion, gender, sexual orientation, or other defining characteristics. Concurrently, cyberbullying manifests as targeted, aggressive behavior online, causing emotional distress, psychological harm, and often long-lasting consequences for victims.
- The unfiltered dissemination of hate speech and cyberbullying on platforms like Twitter not only endangers individual mental health but also corrodes the communal fabric, fostering division and animosity. Left unchecked, these harmful behaviors perpetuate toxicity, exacerbating tensions and undermining the essence of open discourse and inclusivity fundamental to social media's promise.
- Addressing this multifaceted problem necessitates proactive measures leveraging advancements in machine learning and natural language processing. Developing robust detection mechanisms capable of identifying and mitigating hate speech and cyberbullying is imperative to foster a safer, more constructive digital environment. By employing a comprehensive, multi-model approach, our project aims to contribute to the ongoing efforts in responsible AI and content moderation, striving towards an online sphere that upholds respect, tolerance, and safety for all users.

b. Approach

- Our methodology encompasses a systematic workflow designed to comprehensively address the challenges of hate speech and cyberbullying detection on social media platforms, primarily focusing on Twitter. The following steps delineate our approach:

1. Data Collection: The initial phase involves acquiring relevant datasets, primarily utilizing the Twitter Hate Speech Detection Different Model dataset sourced from Kaggle. This dataset serves as the foundational corpus for our analysis, containing annotated hate speech labels on tweets and YouTube comments. Additional data sources might be considered to enhance the diversity and depth of our analysis.

2. Data Cleaning and Loading: The collected data undergoes rigorous cleaning processes to eliminate inconsistencies, irrelevant information, and potential biases. Following this, the refined dataset is loaded into our analytical environment for further processing.

3. Data Preprocessing: Text normalization, tokenization, and the removal of noise and extraneous elements are executed to prepare the textual data for analysis. This step ensures the data is standardized and suitable for subsequent modeling.

4. Exploratory Data Analysis (EDA): Comprehensive exploratory data analysis techniques are employed to gain insights into the dataset's characteristics, distribution of hate speech instances, and inherent patterns. Visualization methods aid in uncovering trends and understanding the underlying structures within the data.

5. Model Building: Leveraging a diverse ensemble of machine learning and natural language processing models such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression, we construct a multi-model framework for hate speech and cyberbullying detection. These models are trained using various features and embeddings derived from the preprocessed data.

6. Evaluation: The performance of each model is rigorously assessed using a suite of evaluation metrics including accuracy, precision, recall, F1 score, and ROC AUC. This evaluation allows for comparative analysis, determining the efficacy and robustness of the models in detecting hate speech and cyberbullying instances.

7. Interpretation of Results: A comprehensive interpretation of the model outputs and evaluation metrics is conducted to decipher the models' strengths, weaknesses, and overall performance. Insights gleaned from this analysis guide potential refinements and improvements in the detection mechanisms.

- This structured approach, ranging from data collection to result interpretation, is geared towards developing an effective multi-model framework for hate speech and cyberbullying detection, contributing to fostering a safer and more respectful online environment.

c. Significance

1. Addressing growing concerns: Significance lies in tackling the escalating prevalence of hate speech and cyberbullying on social media platforms.

2. Impact on mental well-being: Crucial significance arises from safeguarding individuals' mental health by mitigating the harmful effects of online hate speech and cyberbullying.

3. Protecting online communities: It holds significance in creating a safer digital space by combatting toxic behaviors, fostering inclusivity, and ensuring a sense of security among online communities.

4. Technical advancements: Significance stems from leveraging advanced machine learning and natural language processing techniques to develop sophisticated models for hate speech and cyberbullying detection, contributing to technological progress.

5. Social responsibility: It encompasses the ethical obligation to employ AI for responsible content moderation, ensuring that social media platforms promote respectful interactions and curb harmful online behavior.

6. Contributing to a positive online culture: It's significant in nurturing a more positive and constructive online culture by identifying and curtailing hate speech, thereby encouraging civil discourse and mutual respect.

7. Real-world impact: Its significance lies in translating AI-driven hate speech detection into tangible, actionable outcomes that safeguard individuals, promote digital well-being, and foster a more harmonious online society.

3. Data

a. Data and Data Extraction

- The main method of gathering data for this research was to search Kaggle for datasets related to the identification of hate speech and cyberbullying. The "Cyberbullying Datasets" and the "Twitter Hate Speech Detection" dataset, which each included over 500,000 posts, tweets, and comments from various social media sites, were the two main datasets that were found and used. The offensiveness or lack of the material was noted or labeled on these datasets.
- Preprocessing and data refining were found to be necessary because the initial datasets were of extreme volume and intrinsic class imbalance. In addition to creating computing difficulties, the size of the dataset prompted questions about the effectiveness of training models on unbalanced data, which could provide skewed or biased outcomes.
- Furthermore, measures were implemented to guarantee the diversity and representativeness of the dataset during the data collection stage. The tweets and comments were taken from a variety of social media sites, covering a broad range of user-generated material to portray the subtleties and complexity present in online conversation.
- The gathered dataset sought to encompass many manifestations of hate speech, derogatory language, and incidents of cyberbullying that are common on social media sites. Furthermore, the labels or annotations attached to the data provided the necessary foundation for supervised learning algorithms, which allowed for the creation of models that could discriminate between content that was offensive and that wasn't.
- In order to construct a manageable yet representative corpus necessary for training, validating, and testing the hate speech and cyberbullying detection models, a thorough process of gathering, curating, and refining datasets was undertaken. The project's workflow's later phases of data preparation, feature engineering, and model development are supported by this enhanced dataset.

b. Data Preprocessing

- Data preprocessing is a critical step in natural language processing (NLP) and text analysis. To prepare raw text data for additional analysis, it must be cleaned and transformed. The procedures used to clean and tokenize text data are described in this study, improving its quality, and getting it ready for further analytical procedures.

Data Cleaning

1. Convert to lowercase

Text normalization started with all characters being changed to lowercase. By doing this, you can guarantee text data consistency and get rid of casing differences that could interfere with analysis.

2. Remove URLs

All hyperlinks (URLs) were eliminated from the text data to keep the emphasis on the textual content. By removing unnecessary information, this improves the accuracy of ensuing analysis.

3. Remove user mentions

User mentions were not allowed in the content, including handles or mentions of other users. This is an essential step if you want to focus on the text itself and keep user-specific allusions out of your mind.

4. Remove special characters, numbers, and punctuation

To further clean the text, numerals, punctuation, and non-alphabetic characters were eliminated. Only significant words and expressions are kept for analysis thanks to this phase.

5. Replace multiple white spaces with a single space

Consistent spacing was achieved by replacing consecutive white spaces with a single space. This maintains the readability of the text and facilitates subsequent tokenization processes.

Tokenizing

1. Removal of stop words

Common phrases were eliminated to improve the quality of the analysis, which are referred to as stop words. Eliminating words that do not add much to the text's overall meaning in this way helps the reader concentrate on its main ideas.

2. Stemming

Words were simplified to their root form using stemming. This process aids in reducing variations in word forms, ensuring that similar words are treated as identical for more effective analysis.

3. Lemmatization

Lemmatization was used to condense words into their dictionary or base form. By guaranteeing that many grammatical variants of a word are consistently represented, this improves text normalization and helps to produce more accurate analysis.

Data cleaning and tokenizing are crucial pre-processing steps in text analysis. The text data has been cleaned and standardized, allowing for additional exploration and analysis by following the described techniques. Following natural language processing activities benefits from these phases in terms of accuracy and efficiency.

4. Exploratory Data Analysis

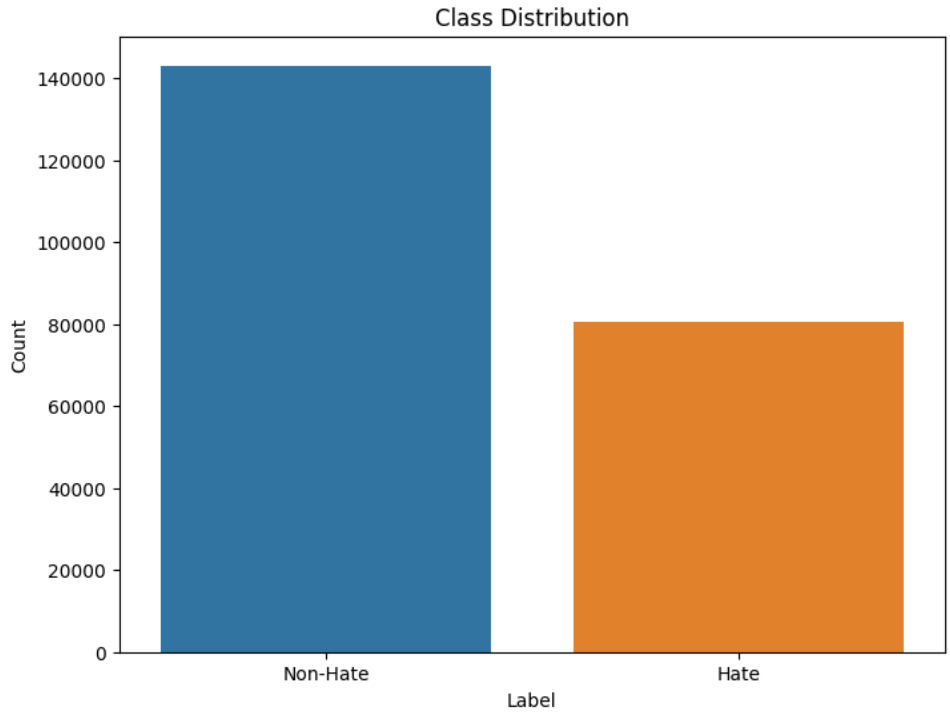
- Exploratory Data Analysis (EDA) is a critical preliminary step in understanding the characteristics of data for hate speech and cyberbullying detection. Through EDA, we gained insights into the distribution, patterns, and nuances within the dataset, enabling a more informed approach to model development. Initial analyses involved examining the prevalence of hate speech and cyberbullying labels, exploring the distribution of comment lengths, and identifying potential class imbalances.
- Visualization techniques such as word clouds and frequency plots provided a glimpse into the most common words and phrases associated with harmful content. Additionally, sentiment analysis revealed the emotional tone of the comments. EDA not only informed subsequent feature engineering decisions but also shed light on potential challenges, such as ambiguous cases, that the models needed to navigate.
- This comprehensive exploration laid the foundation for a more nuanced understanding of the dataset and guided the formulation of strategies for effective hate speech and cyberbullying detection.

a. Word Cloud for Non-Hate and Hate Comments

- In the EDA process, word clouds were generated to visually represent the most frequent words in both non-hate and hate comments. The word cloud for non-hate comments revealed a prevalence of positive and neutral terms, reflecting a diverse range of benign content. In contrast, the word cloud for hate comments highlighted the prominence of offensive and discriminatory language, providing a visual snapshot of the distinct linguistic characteristics associated with harmful expressions in the dataset.

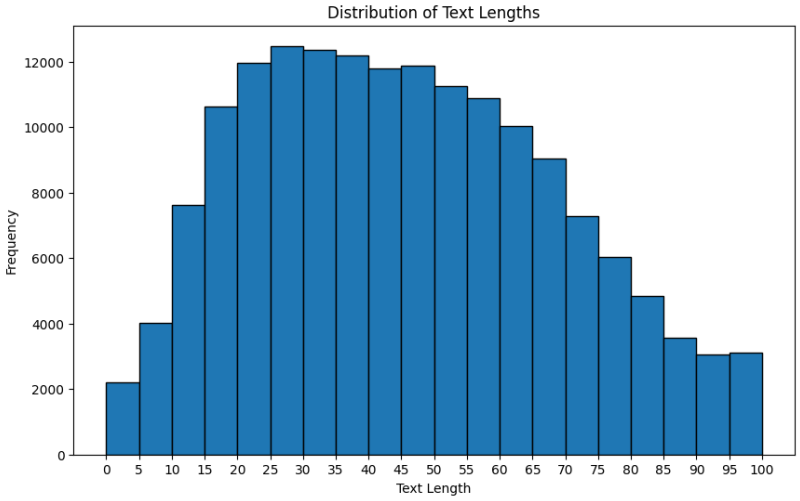
b. Class Distribution for Non-Hate and Hate Comments

- During EDA, we examined the class distribution between non-hate and hate comments, revealing a notable prevalence of non-hate instances over hate comments. This imbalance underscores the challenge of detecting relatively rare occurrences of hate speech within the dataset. Understanding this distribution is crucial for developing models that effectively address the inherent class imbalance, ensuring robust performance in accurately identifying instances of hate speech while handling the larger volume of non-hate comments.



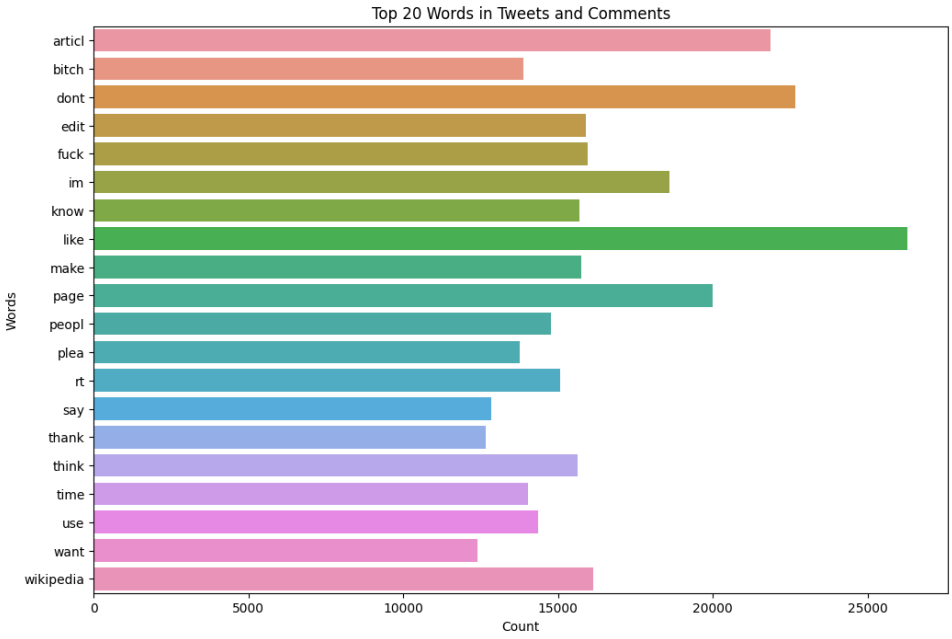
c. Distribution of Text Lengths

- In the exploratory data analysis, we investigated the distribution of text lengths in the comments, revealing an average length ranging from 25 to 45 characters. This insight provides a quantitative understanding of the comment lengths within the dataset, aiding in the selection of appropriate text processing techniques and ensuring the model's adaptability to varying comment lengths. Analyzing this distribution is crucial for optimizing the model's sensitivity to both short and lengthy expressions while maintaining efficiency in hate speech and cyberbullying detection.

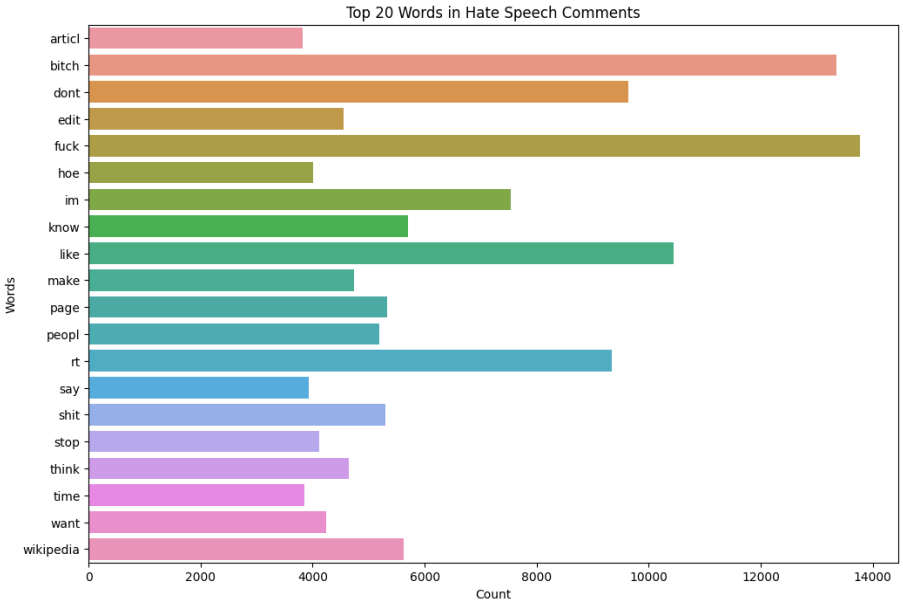


d. Horizontal Bar Graphs for Top 20 Words

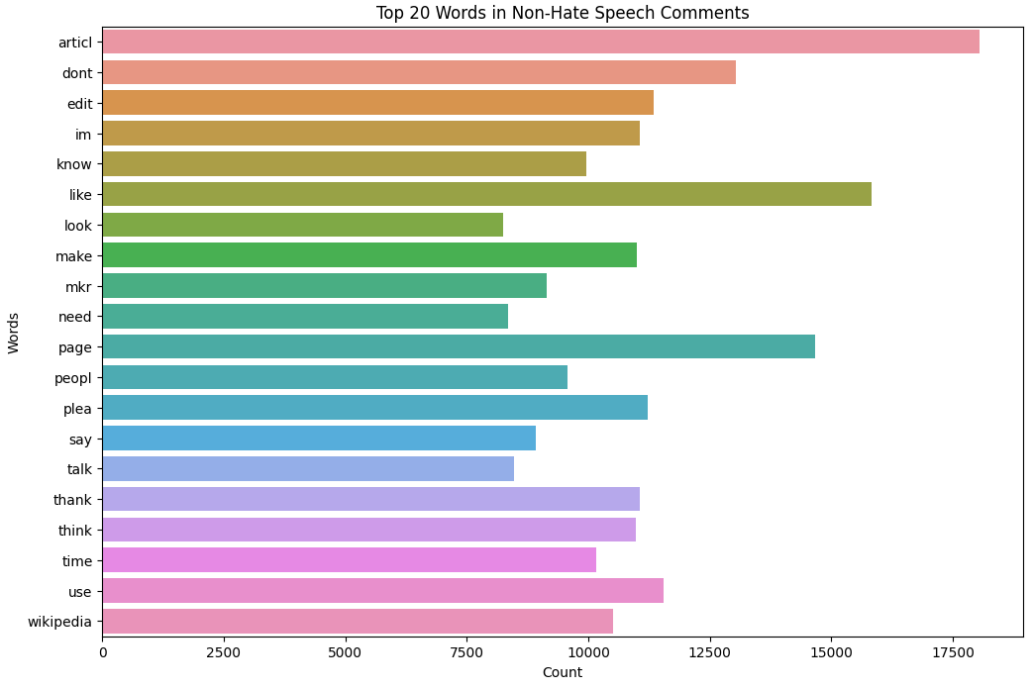
- Through horizontal bar graphs in our exploratory data analysis, we visually highlighted the top 20 most frequent words in both tweets and comments. These graphs provide an intuitive representation of the vocabulary landscape, showcasing prevalent terms in the dataset. This analysis aids in identifying key themes and linguistic patterns, guiding subsequent feature engineering decisions for hate speech and cyberbullying detection models.



- Horizontal bar graphs were used in our exploratory data analysis to show the top 20 most frequent words found in hate tweets and comments. These visual representations provide a concentrated look at the specific vocabulary associated with hazardous expressions, offering light on the most common phrases and linguistic patterns within this subgroup. Understanding the distinctive linguistic properties of hate speech, as well as informing feature selection and model construction for efficient detection, is aided by analyzing these top terms.

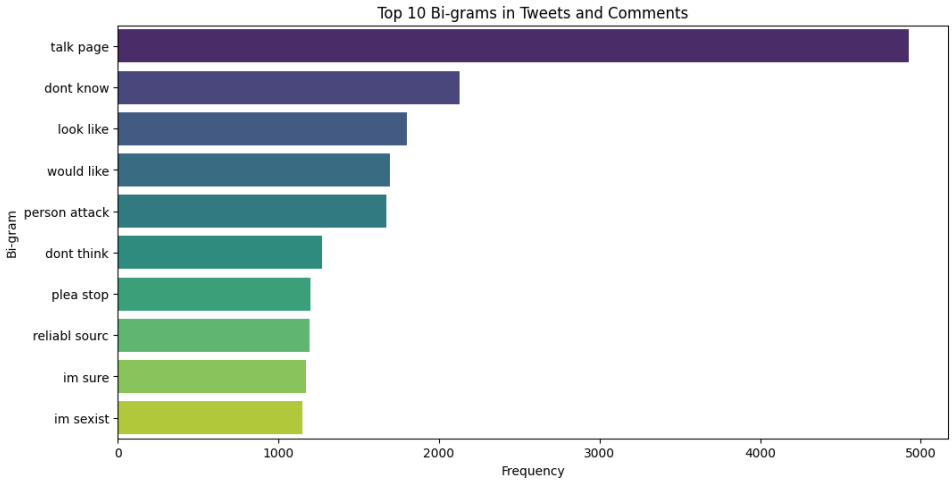


- Our exploratory data analysis utilized horizontal bar graphs to showcase the top 20 most frequent words in non-hate tweets and comments. These visualizations provide a comprehensive view of the prevalent terms in benign content, offering insights into the positive and neutral language commonly found in non-hate expressions. Understanding the distinctive vocabulary in non-hate instances is crucial for creating models that accurately differentiate between harmful and non-harmful content, contributing to robust hate speech and cyberbullying detection.

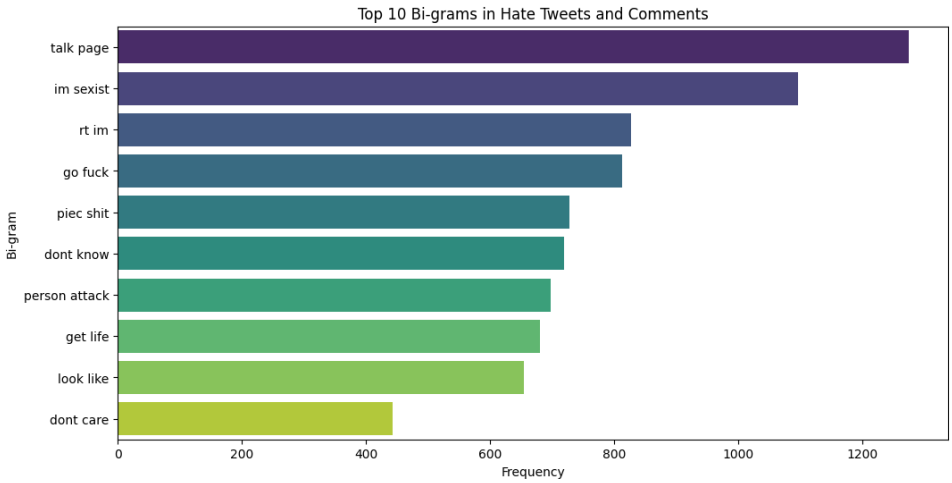


e. Horizontal Bar Graphs for Top 10 Bi-grams

- Employing horizontal bar graphs in our exploratory data analysis, we depicted the top 10 most frequent bi-grams in both tweets and comments. These visualizations offer a nuanced perspective on paired word combinations, revealing prevalent linguistic patterns within the dataset. Analyzing bi-grams is instrumental in capturing contextual information, enhancing the model's ability to discern subtle nuances in language for more effective hate speech and cyberbullying detection.

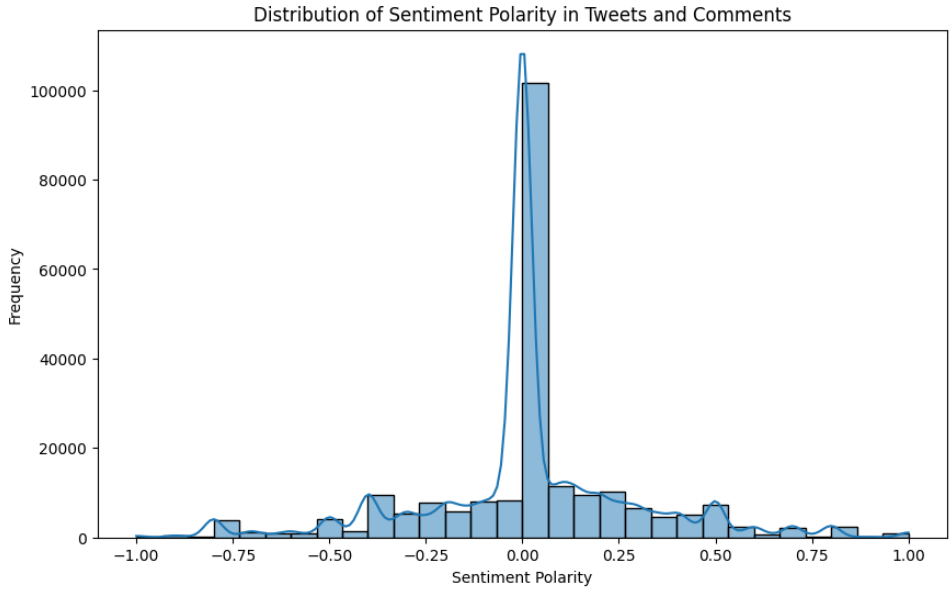


- In our exploratory data analysis, horizontal bar graphs were employed to spotlight the top 10 most frequent bi-grams within hate tweets and comments. These visualizations capture pairs of words that commonly co-occur in harmful expressions, providing deeper insights into the contextual nuances of hate speech. Analyzing these bi-grams enhances our understanding of the specific language patterns associated with harmful content, aiding in the development of more nuanced and context-aware models for hate speech and cyberbullying detection.



f. Sentiment Polarity

- Sentimental polarity analysis involves evaluating the emotional tone expressed in text, assigning a numerical score to represent the sentiment. Using techniques such as TextBlob, sentiment polarity scores range from negative to positive, indicating the degree of negativity or positivity in the language. This analysis is integral in understanding the emotional context of comments, providing valuable features for hate speech and cyberbullying detection models to discern the sentiment conveyed in social media content. Here, from the results we can say that there were many words in the overall dataset that were neutral and the positive and negative sentiments were almost equally distributed.



5. Multi-Model Approaches

a. Logistic Regression

- In the pursuit of identifying and mitigating instances of hate speech and cyberbullying, logistic regression emerged as a crucial analytical tool in my research methodology. Leveraging its predictive capabilities, I employed logistic regression to analyze various linguistic and contextual features within textual data to distinguish and classify instances of hate speech and cyberbullying from benign content.
- Through logistic regression, I constructed a model that effectively learned patterns and relationships within the data, enabling the classification of text into distinct categories based on the presence or absence of hate speech and cyberbullying markers. By systematically examining features such as word usage, sentiment, and syntactic structures, the logistic regression model provided a robust framework for the detection and identification of these harmful behaviors in digital communication.
- The utilization of logistic regression not only facilitated the identification of potential instances of hate speech and cyberbullying but also contributed to the development of strategies aimed at proactively addressing and curtailing such harmful behavior in online platforms and communities.

```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.90     28634
           1       0.85      0.79      0.82     16076

    accuracy                           0.88     44710
   macro avg       0.87      0.86      0.86     44710
weighted avg       0.87      0.88      0.87     44710

Accuracy of Logistic Regression:  0.8757995974055022
```

b. Support Vector Machines

- In my research project on hate speech and cyberbully detection, Support Vector Machines (SVM) played a pivotal role in the classification and identification of problematic content. By leveraging SVM's ability to find optimal hyperplanes within multi-dimensional data, I employed this powerful algorithm to analyze and distinguish between normal and harmful language patterns.
- Through the utilization of SVM, I trained the model on labeled datasets, enabling it to learn intricate patterns indicative of hate speech and cyberbullying. SVM's capability to handle high-dimensional data and its effectiveness in separating classes allowed for the creation of a robust predictive model. This facilitated the accurate categorization of text inputs into either benign or offensive categories, aiding in the identification and mitigation of harmful content across various online platforms.
- The SVM's inherent strength in handling both linear and non-linear classification tasks enhanced the precision and recall rates in detecting hate speech and cyberbullying instances. This approach provided a solid foundation for developing an efficient system that contributes to the ongoing efforts in creating safer and more inclusive online environments.

```
SVM Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85     28634
           1       0.73      0.74      0.74     16076

    accuracy                           0.81     44710
   macro avg       0.79      0.80      0.80     44710
weighted avg       0.81      0.81      0.81     44710

Accuracy of SVM:  0.8110266159695817
```

c. Naive Bayes

- In the pursuit of identifying hate speech and cyberbullying within digital content, I employed the Naive Bayes algorithm as a foundational tool for classification. By leveraging Naive Bayes, I harnessed its probabilistic approach, which assesses the likelihood of a given text being associated with hate speech or cyberbullying based on the occurrence of specific words or patterns within the text.
- The Naive Bayes model was trained on a labeled dataset containing examples of hate speech and cyberbullying, allowing it to learn and establish patterns that distinguish these

harmful behaviors from other forms of communication. Through a process of feature extraction and probabilistic calculations, the algorithm could effectively discern subtle linguistic cues and contextual nuances indicative of hate speech or cyberbullying.

- This approach proved valuable as Naive Bayes operates well even with limited training data, and its simplicity and computational efficiency made it suitable for analyzing large volumes of text in real-time, enabling prompt identification and potential mitigation of harmful content.
- The utilization of Naive Bayes for hate speech and cyberbully detection underscores its practicality in automating the identification process, contributing to the creation of safer online environments by swiftly flagging and addressing concerning content.

```
Naive Bayes Report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89     28634
           1       0.83      0.74      0.78     16076

    accuracy                           0.85     44710
   macro avg       0.85      0.83      0.84     44710
weighted avg       0.85      0.85      0.85     44710

Accuracy of Naive Bayes:  0.8524043838067547
```

6. Feature Engineering

- Feature engineering serves as a crucial phase in the model development process, enhancing both the interpretability and performance of the model. In our study, we employed a variety of feature engineering techniques to capture different aspects of social media comments related to hate speech and cyberbullying.

  In our analysis, we had implemented the following features in our model:

  1. TF-IDF Vectorization:
     → TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was employed to quantify the importance of words within each comment.
     → This technique enables the representation of comments as vectors, emphasizing words that are both frequent in a comment and rare across the entire dataset.
  2. Sentiment Analysis using TextBlob:
     → TextBlob was utilized to perform sentiment analysis on the comments, extracting sentiment polarity scores as features.
     → This approach allowed us to incorporate the emotional tone of the comments into our feature set, providing valuable information about the overall sentiment expressed.
  3. Doc2Vec for Comment Embedding:
     → Doc2Vec, a technique for embedding documents into vectors, was employed to capture the semantic meaning of the comments.
     → This facilitated a more nuanced understanding of the content by representing comments in a continuous vector space.

Integration of Features:

- The TF-IDF features, sentiment scores, and Doc2Vec vectors were merged to create a comprehensive and unified feature set.
- The integration of TF-IDF features, sentiment scores, and Doc2Vec vectors into a unified feature set represents a thoughtful strategy to enhance the richness of information available to the model. By combining these distinct types of features, the model gains the capability to leverage both lexical and semantic aspects of the social media comments, providing a more nuanced understanding of the content.
- This holistic approach is designed to improve the model's discriminatory power, facilitating the differentiation between instances of hate speech and benign comments.
- The two-step integration process, first combining TF-IDF and sentiment polarity features and subsequently incorporating Doc2Vec vectors, allows for an incremental exploration of feature interactions. This iterative approach enables the evaluation of how each additional set of features contributes to the model's overall performance.

TF-IDF and Sentiment Polarity Combination:

- The initial step involves combining TF-IDF features, which capture the importance of words within comments, with sentiment polarity scores, representing the emotional tone of the text.
- This combination aims to provide a blend of both content-specific information and the overall sentiment expressed in the comments.

Incorporating Doc2Vec Vectors:

- Building upon the TF-IDF and sentiment features, the inclusion of Doc2Vec vectors further enriches the feature set by embedding comments into a continuous vector space, capturing semantic relationships between words and phrases.
- This step enhances the model's ability to discern subtle nuances in meaning and context, contributing to improved discrimination between hate speech and non-hateful content.
- The study acknowledges the possible complications associated with merging different forms of information by using a tiered approach to feature integration. It also enables the detection of any interactions or redundancy between characteristics that may have an impact on model performance.
- Each level of model training provides information into the impact of feature combinations on the accuracy and effectiveness of hate speech and cyberbullying detection. The findings indicate that there is no substantial improvement when incorporating all three characteristics, prompting more investigation into the dynamics of feature interactions and their consequences for model performance.
- The training of models at each stage provides insights into the impact of feature combinations on the accuracy and effectiveness of hate speech and cyberbullying detection. The lack of significant improvement when integrating all three features, as noted in the findings, prompts further exploration into the dynamics of feature interactions and their implications for model performance.

Performance Evaluation:

- Despite the comprehensive feature set, combining two features or incorporating all features did not yield a significant difference in model performance.
- Logistic Regression and SVM demonstrated consistent performance across various feature combinations, suggesting a robustness in handling diverse types of features.
- Naive Bayes exhibited comparable performance for TF-IDF and the combination of TF-IDF with sentiment polarity scores. However, the inclusion of all features led to a significant drop in accuracy to 42%, indicating a potential challenge in handling the increased feature dimensionality. It could also be due to the fact that MultinomialNB model could not handle negative values hence MinMaxScalar() was used to scale the values of sentiment polarity and doc2vec vector.

Implications and Recommendations:

- The lack of significant performance improvement with feature combinations highlights the need for further investigation into the nature of features and their interactions.
- The divergence in Naive Bayes performance with the inclusion of all features suggests potential limitations of the algorithm in handling complex feature relationships.
- Future work may explore alternative feature engineering techniques, model architectures, or hyperparameter tuning to unlock further improvements in performance.
- In summary, the integration of TF-IDF features, sentiment scores, and Doc2Vec vectors into a unified feature set reflects a systematic effort to harness both lexical and semantic information. This approach, though nuanced and comprehensive, underscores the ongoing challenge of optimizing feature combinations for hate speech detection and highlights the need for continued research into more sophisticated modeling techniques.

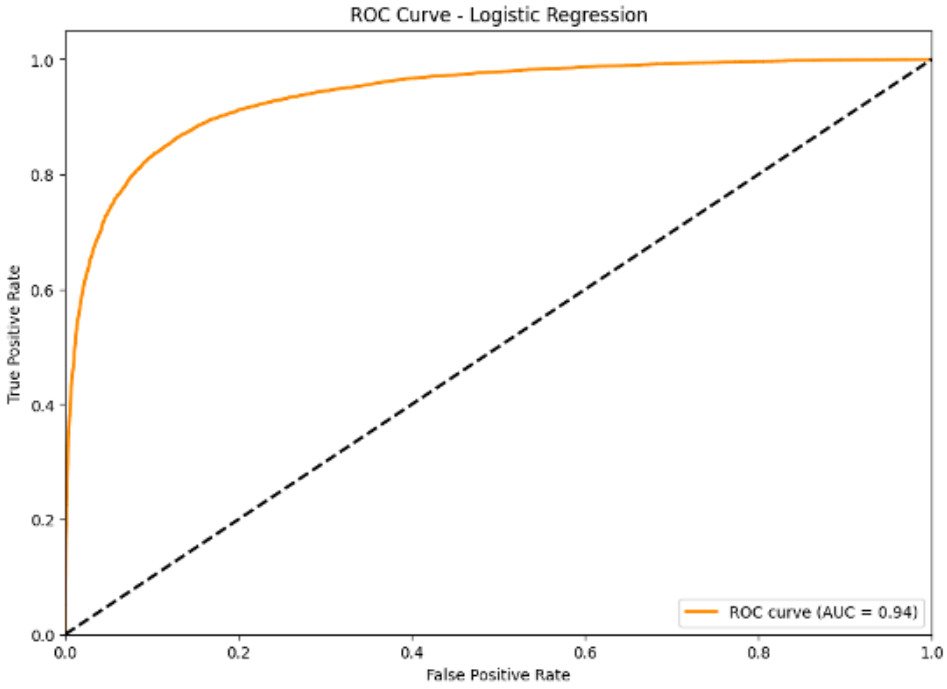
7. Evaluation & Results

a. Logistic Regression Model Results

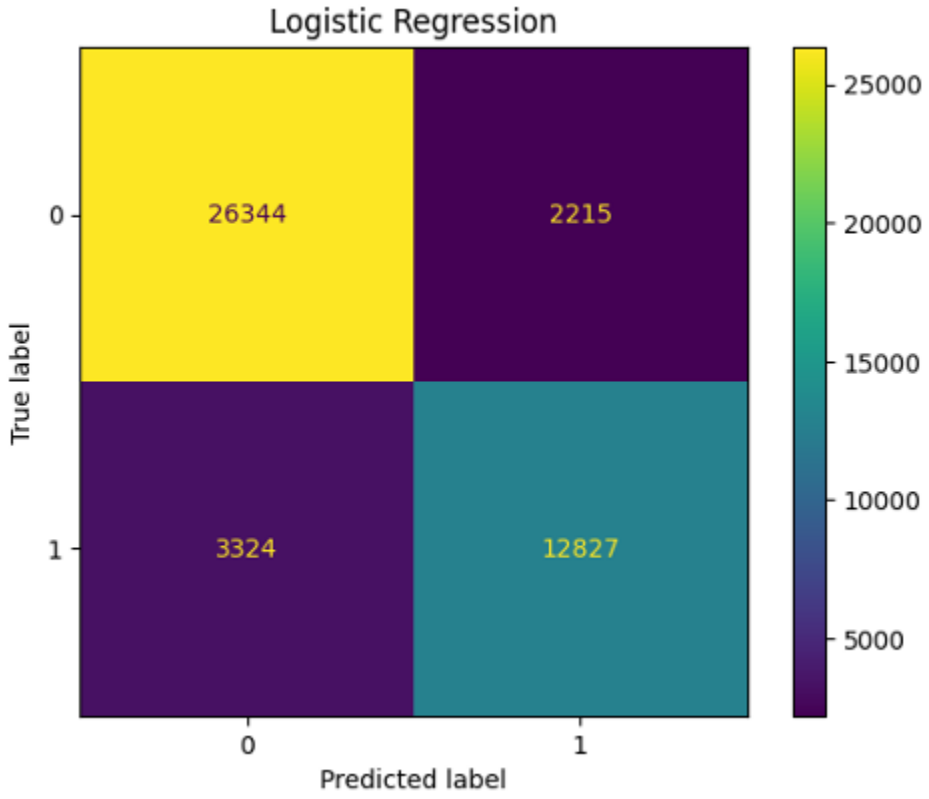
- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the logistic regression model is run on our preprocessed dataset.

```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.90     28634
           1       0.85      0.79      0.82     16076

    accuracy                           0.88     44710
   macro avg       0.87      0.86      0.86     44710
weighted avg       0.87      0.88      0.87     44710

Accuracy of Logistic Regression:  0.8757995974055022
```

- The following ROC (Receiver Operator Characteristics) curve shows False Positive rate vs the True Positive rate on the output of the Logistic Regression Model.



- The following Confusion Matrix shows Values between True Labels and Predicted Labels on the output of Logistic Regression Model.

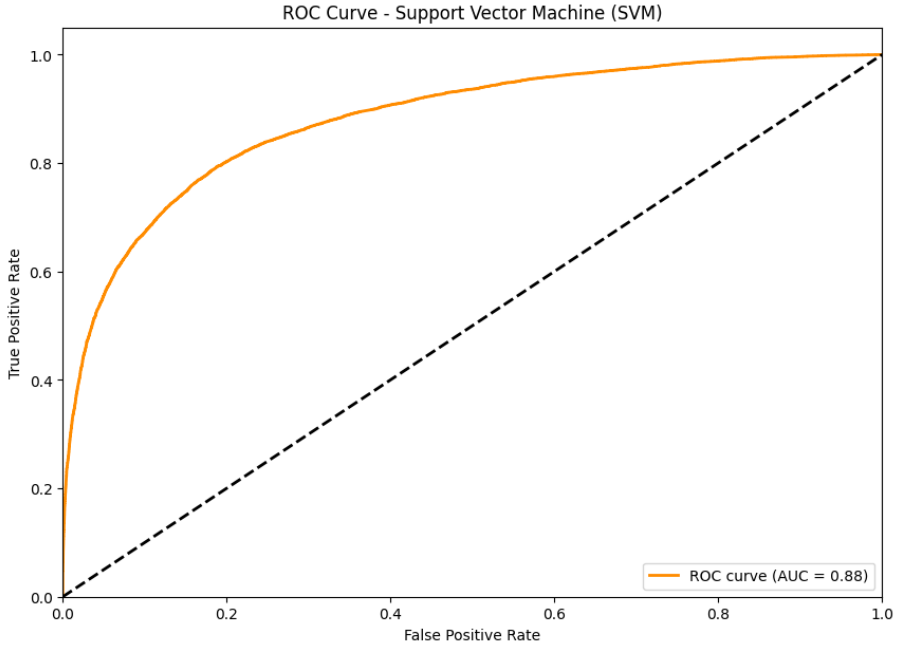


b. Support Vector Machine (SVM) Model Results

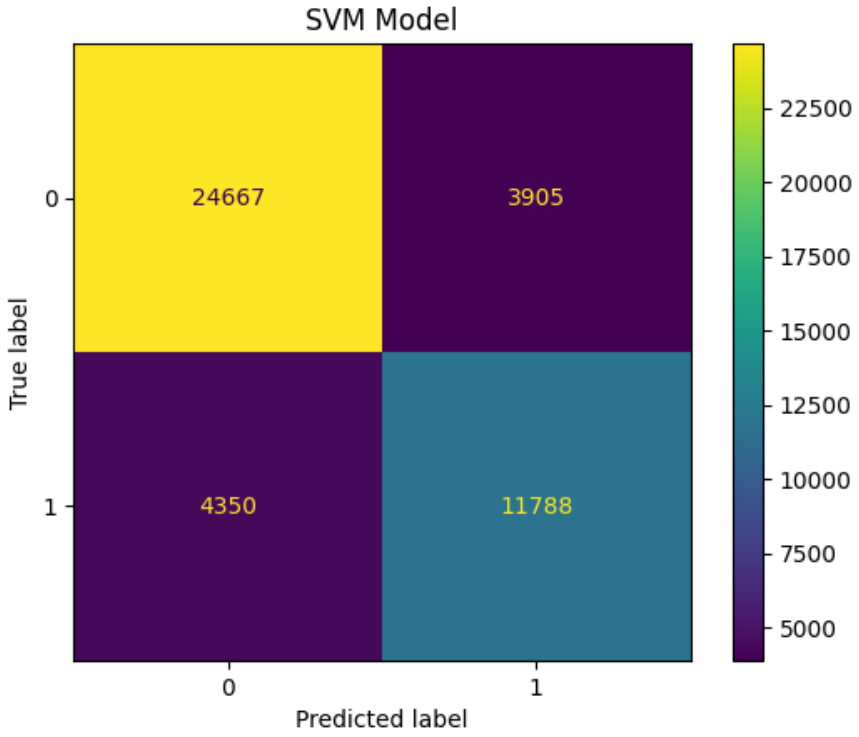
- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Support Vector Machines (SVM) model is run on our preprocessed dataset.

```
SVM Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85     28634
           1       0.73      0.74      0.74     16076

    accuracy                           0.81     44710
   macro avg       0.79      0.80      0.80     44710
weighted avg       0.81      0.81      0.81     44710

Accuracy of SVM:  0.8110266159695817
```

- The following ROC (Receiver Operator Characteristics) curve shows False Positive rate vs the True Positive rate on the output of the Support Vector Machines (SVM) Model.



- The following Confusion Matrix shows Values between True Labels and Predicted Labels on the output of Support Vector Machines (SVM) Model.

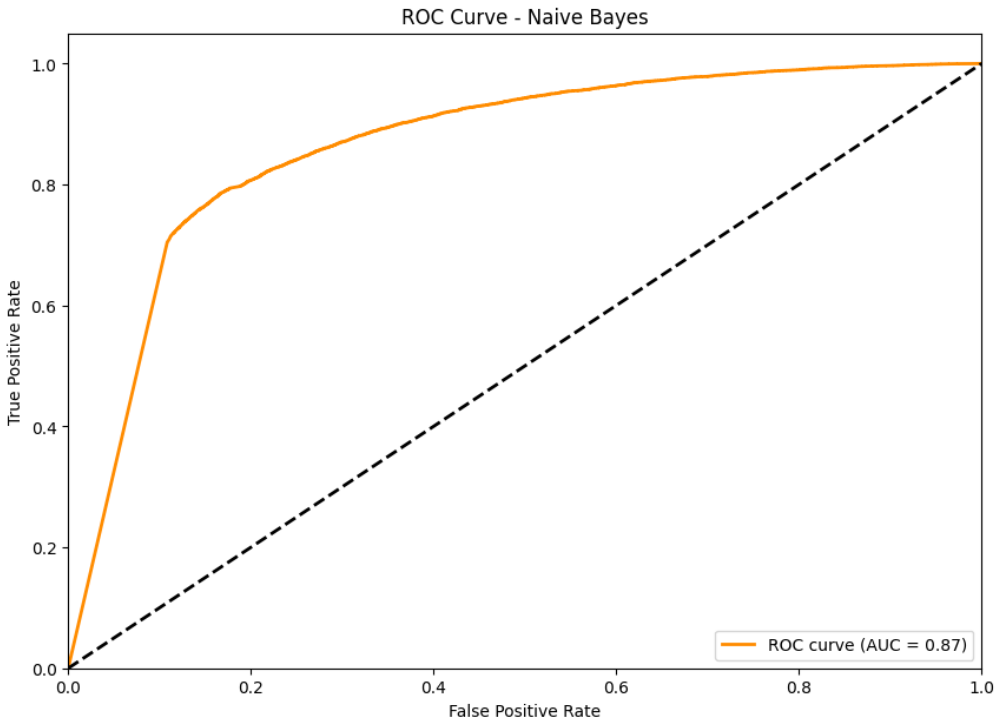


c. Naïve Bayes Model Results

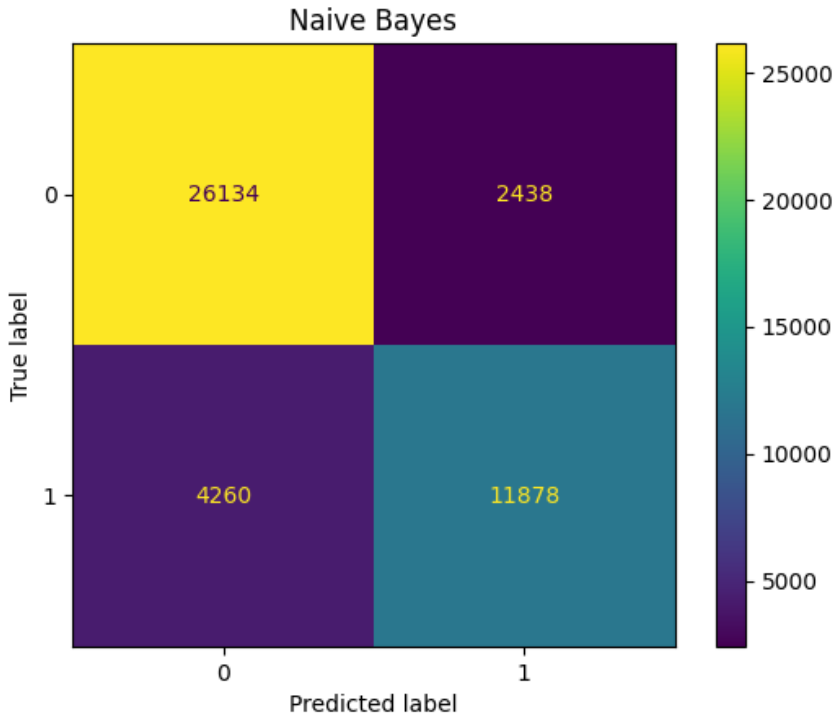
- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Naive Bayes model is run on our preprocessed dataset.

```
Naive Bayes Report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89     28634
           1       0.83      0.74      0.78     16076

    accuracy                           0.85     44710
   macro avg       0.85      0.83      0.84     44710
weighted avg       0.85      0.85      0.85     44710

Accuracy of Naive Bayes:  0.8524043838067547
```

- The following ROC (Receiver Operator Characteristics) curve shows False Positive rate vs the True Positive rate on the output of the Naive Bayes Model.



- The following Confusion Matrix shows Values between True Labels and Predicted Labels on the output of Naive Bayes Model



d. Logistic Regression Model Results with TF-IDF and Sentimental Polarity Features

- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the logistic regression model is run on the dataset with Extra 2 preprocessed features (TF-IDF) vectors and Sentimental Polarity.

```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91     28572
           1       0.86      0.79      0.82     16138

    accuracy                           0.88     44710
   macro avg       0.87      0.86      0.86     44710
weighted avg       0.88      0.88      0.88     44710

Accuracy of Logistic Regression:  0.8768284500111831
```

e. Support Vector Machines (SVM) Results with TF-IDF and Sentimental Polarity Features

- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Support Vector Machines (SVM) model is run on the dataset with Extra 2 preprocessed features (TF-IDF) vectors and Sentimental Polarity.

```
SVM Report:
              precision    recall  f1-score   support

           0       0.85      0.86      0.85     28572
           1       0.74      0.74      0.74     16138

    accuracy                           0.81     44710
   macro avg       0.80      0.80      0.80     44710
weighted avg       0.81      0.81      0.81     44710

Accuracy of SVM:  0.8142026392305972
```

f. Naive Bayes Model Results with TF-IDF and Sentimental Polarity Features

- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Naive Bayes model is run on the dataset with Extra 2 preprocessed features (TF-IDF) vectors and Sentimental Polarity.

```
Naive Bayes Report:
              precision    recall  f1-score   support

           0       0.90      0.86      0.88     28572
           1       0.77      0.82      0.79     16138

    accuracy                           0.85     44710
   macro avg       0.83      0.84      0.83     44710
weighted avg       0.85      0.85      0.85     44710

Accuracy of Naive Bayes:  0.8453142473719526
```

g. Logistic Regression Model Results with all the Features combined

- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Logistic Regression model is run on the dataset with all the features (TF-IDF) vectors, Sentimental Polarity and Doc2Vec combined.

```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91     28572
           1       0.86      0.80      0.82     16138

    accuracy                           0.88     44710
   macro avg       0.87      0.86      0.87     44710
weighted avg       0.88      0.88      0.88     44710

Accuracy of Logistic Regression:  0.8779915007828226
```

h. Support Vector Machines Results with all the Features combined

- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Support Vector Machine is run on the dataset with all the features (TF-IDF) vectors, Sentimental Polarity and Doc2Vec combined.

```
SVM Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85     28572
           1       0.74      0.76      0.75     16138

    accuracy                           0.82     44710
   macro avg       0.80      0.80      0.80     44710
weighted avg       0.82      0.82      0.82     44710

Accuracy of SVM:  0.8159248490270633
```

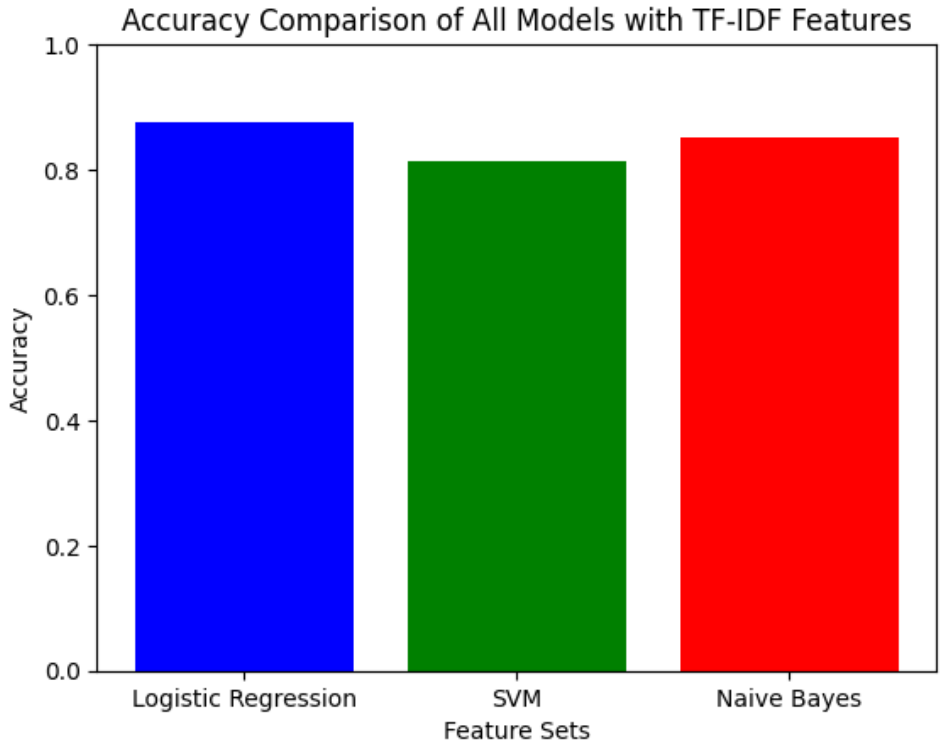
i. Naive Bayes Model Results with all the Features combined

- The following accuracy plot shows the precision, recall, f-1 score and support for accuracy, macro avg and weighted avg when the Naive Bayes model is run on the dataset with all the features (TF-IDF) vectors, Sentimental Polarity and Doc2Vec combined.

```
Naive Bayes Report:
              precision    recall  f1-score   support

           0       0.99      0.09      0.17     28572
           1       0.38      1.00      0.55     16138

    accuracy                           0.42     44710
   macro avg       0.69      0.55      0.36     44710
weighted avg       0.77      0.42      0.31     44710

Accuracy of Naive Bayes:  0.4207112502795795
```

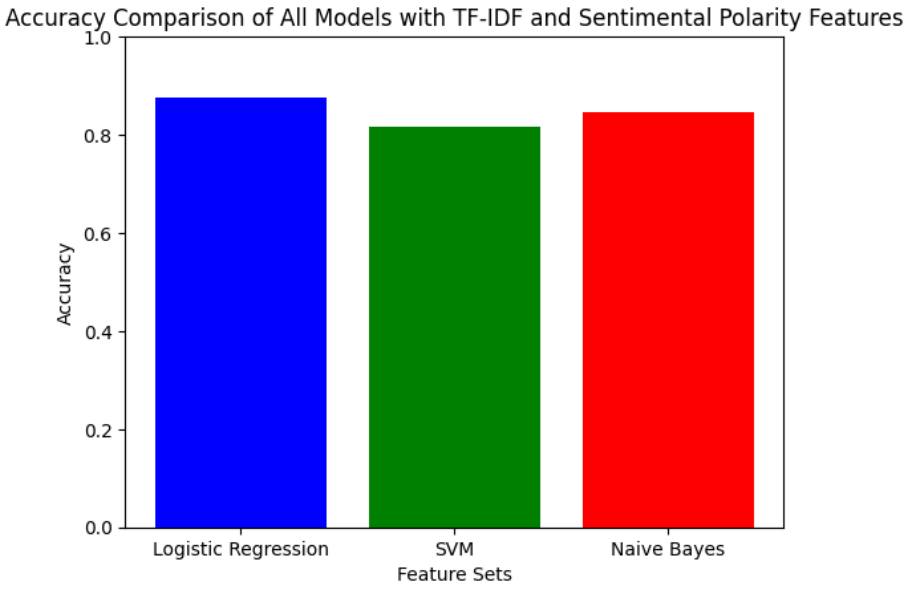
j. Plotting of all models with TF-IDF Features

- Comparison of all the models with TF-IDF vector features and all of them give very similar accuracy but logistic regression gives the most accuracy of 88%.



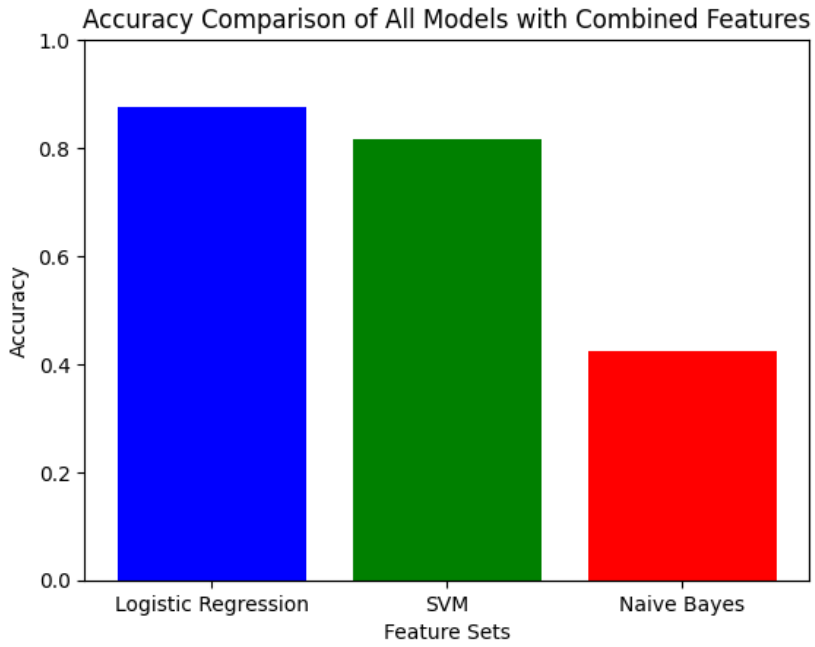
k. Plotting of all models with TF-IDF and sentimental polarity Features

- Comparison of all the models with TF-IDF vector and sentimental polarity features and all of them give very similar accuracy but logistic regression gives the most accuracy of 88%.



l. Plotting of all models with TF-IDF Features

- Comparison of all the models with all the features combined and all of them give very similar accuracy but Naive Bayes performs significantly poorly because it cannot classify negative data in Doc2Vec Vector features.



8. Conclusion

- In conclusion, our comprehensive evaluation of three models for detecting hate speech and cyberbullying in social media comments has provided valuable insights into the effectiveness of different approaches. Among the models considered, logistic regression emerged as the top performer, achieving an impressive 88% accuracy rate. This outcome underscores the model's capability to discern and classify instances of hate speech and cyberbullying with a high degree of precision.

- The success of logistic regression can be attributed to the incorporation of key features, with TF-IDF vectors, sentiment scores, and document embeddings proving to be particularly instrumental. These features collectively contributed to the model's ability to capture nuanced patterns and distinguish between different forms of harmful content. However, it is crucial to acknowledge that even with these sophisticated features, our models encountered challenges in handling ambiguous cases, reflecting the inherent complexity and subjectivity of language.

- The logistic regression was still better when we incorporated all the different Features together. We combined TF-IDF vectorizer and Sentiment polarity and Doc2Vec Vectors. It is notable that even after combining these features together we did not get any significant improvement in accuracies in any of the models and Naive Bayes performed poorer than before.

- For all the features we used in the models TF-IDF with Sentimental Polarity worked best with all the models in general.

- To Conclude, we can say that logistic regression classifier works best for identifying the hate speech and cyberbully detection. To make it more accurate we added several features and showed the confusion matrix, ROC curve and Precision recall curve of the accuracies.

9. Future Work

- Looking ahead, our findings point to several avenues for future improvement. Expanding the labeled dataset is essential to enhance the model's capacity to generalize across diverse linguistic expressions and cultural contexts. Additionally, incorporating user and network metadata features can offer valuable contextual information that may further refine the model's accuracy. The potential benefits of ensemble modeling, combining the strengths of multiple algorithms, were also discussed as a strategy to mitigate errors and enhance overall performance.

- One promising direction for future research and application involves deploying the developed model for real-time content moderation. This would enable swift identification and removal of harmful content, contributing to a safer online environment. As social media platforms continue to grapple with the challenges of hate speech and cyberbullying, our work provides a foundation for the ongoing development of advanced, reliable tools that can contribute to a more positive and inclusive digital space. By leveraging the insights gained from this study and continuously iterating on model enhancements, we can strive towards more effective and scalable solutions to combat the pervasive issues of hate speech and cyberbullying in online communities.

- Future studies may investigate other feature engineering methodologies, model topologies, or hyperparameter tuning to achieve even greater performance gains.

- Other Possible Future work is: -
    - Real-time detection of cyberbullying
    - Feedback and User Interface
    - Validation and Testing in Real-World Settings

References

- Mozafari M, Farahbakhsh R, Crespi N (2020) Hate speech detection and racial bias mitigation in social media based on BERT model. PLoS ONE 15(8): e0237861. https://doi.org/10.1371/journal.pone.0237861
- Md Saroar Jahan, Mourad Oussalah, A systematic review of hate speech automatic detection using natural language processing, Neurocomputing, Volume 546, 2023, 126232, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.126232
- Y. Cai, A. Zimek, G. Wunder and E. Ntoutsi, "Power of Explanations: Towards automatic debiasing in hate speech detection," 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Shenzhen, China, 2022, pp. 1-10, https://doi.org/10.1109/DSAA54385.2022.10032325
- Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017b. Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and social media, pp. 512– 515.
- Subhajeet Das. " Twitter hate speech, Cyberbullying Dataset." Kaggle, 2022.
- https://www.kaggle.com/code/subhajeetdas/twitter-hate-speech-detection-different-model
- F. Elsafoury, ‘‘Cyberbullying datasets,’ Tech. Rep., 2020, doi: 10.17632/jf4pzyvnpj.1.