CSE422: ARTIFICIAL INTELLIGENCE

Report: Sentiment Analysis on Twitter Data

Group - 3

Submitted by:
Shreya Adrita Banik- 21301669
Niaz Nafi Rahman 21301700
Tahsina Moiukh- 21301612
Zaid Rehman- 21301389

1. Introduction:

The project's primary aim is to develop a model capable of accurately classifying positive and negative sentiments in tweets. This is accomplished by employing a logistic regression model and a Naive Bayes Model to analyze the sentiment expressed in each tweet and categorize it accordingly.

The primary challenge that this project aims to address is the widespread occurrence of hate speech on social media sites like Twitter, especially remarks that are sexist and racist. These kinds of statements may result in a hostile online environment and have a profoundly detrimental influence on individuals as well as communities.

This project has several different motivations:

1. **Social responsibility:** As social media has grown in popularity, it is more important than ever to keep an eye on and restrict the distribution of damaging content. Through the automatic detection of hate speech, this project helps to make the internet a more secure and welcoming place.
2. **Technological Advancement:** By demonstrating how these technologies may be used to address real-world issues, this initiative also seeks to further the fields of machine learning and natural language processing (NLP).
3. **Research Contribution:** By addressing a challenging and important societal issue, the project offers the research community useful methods and insights, especially in the fields of automated content moderation and sentiment analysis.

It is anticipated that upon completion of this project, a model using logistic regression and another using Naive Bayes will be developed

that can accurately distinguish between tweets that are racist or sexist and those that are not. To ensure a fair assessment of this model's recall and precision, its performance will be evaluated using the F1 score and accuracy.

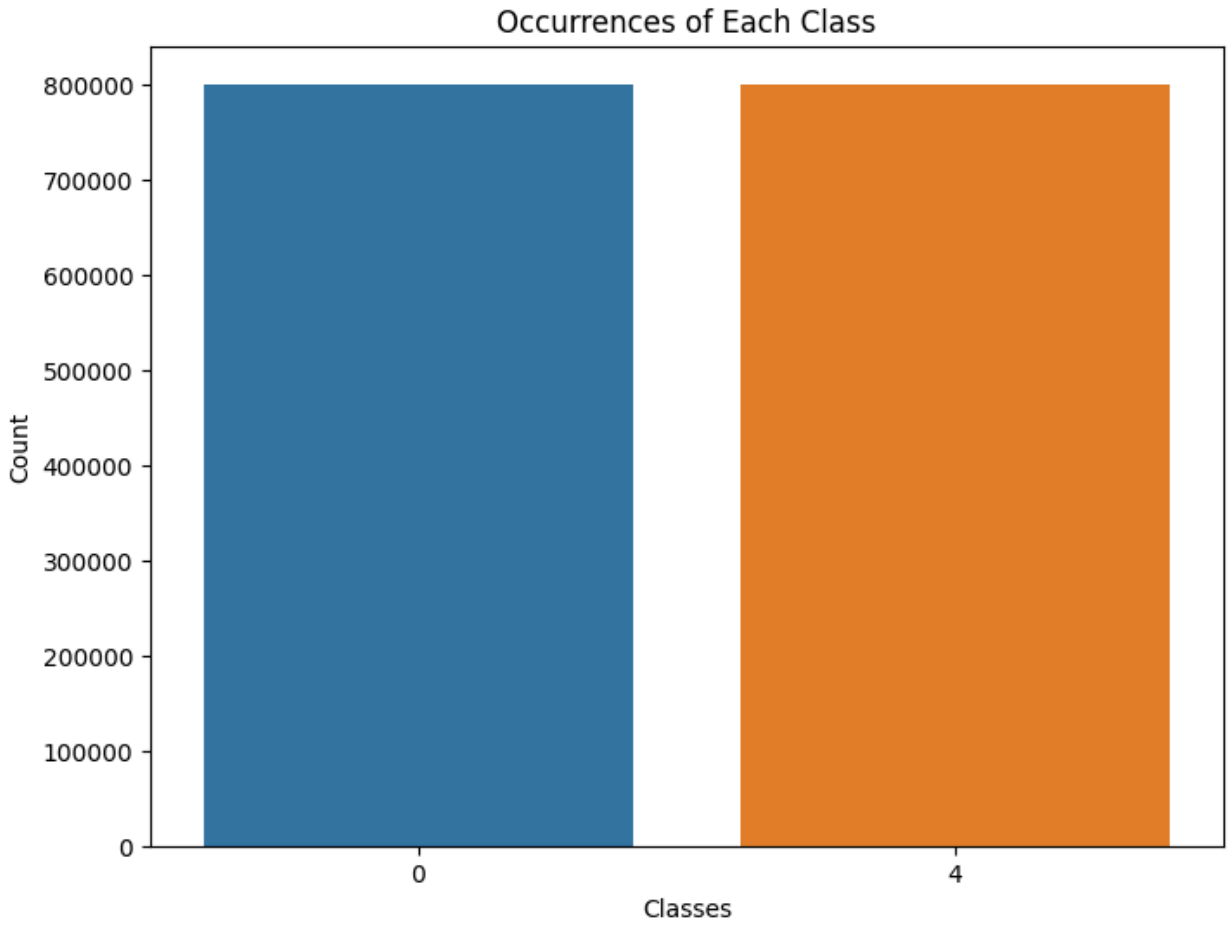
In conclusion, our project seeks to develop machine learning and natural language processing capabilities alongside using technology to address an urgent societal issue. Through this, it has the potential of having an enormous impact on promoting respectful and inclusive communication in both the digital and larger contexts.

2. Dataset Description:

- **Source:** https://www.kaggle.com/datasets/kazanova/sentiment140/data
- **Reference:** https://web.stanford.edu/~jurafsky/slp3/
- **Features:** The dataset contains six features including 'id', 'date', 'query', 'user', 'tweet' and label
- **Problem Type: Classification**
    - Justification: The sentiment analysis task involves classifying tweets into positive (4, later replaced to 1) or negative (0), making it a binary classification problem.
- **Data Points:** 1,600,000
- **Features:** The dataset includes categorical features. The sentiment analysis task primarily involves text data ('tweet' column) and the sentiment label ('label') after preprocessing. In this context:
    - **Tweet Text ('tweet' column):** This categorical feature represents the textual content of tweets. Preprocessing steps such as cleaning, normalization, and stemming were applied to this feature to refine and standardize the text data for analysis. The resulting cleaned and processed text was used for feature extraction via the Bag-of-Words (BoW) technique.
    - **Sentiment Label ('label' column):** Initially ranging from 0 to 4, this categorical feature underwent label transformation to convert

sentiment classes into a binary classification—0 for negative sentiment and 1 for positive sentiment.

- **Feature Correlation:** In context of our text classification problem using bag-of-words representation, the features are word counts or binary counts for the indication of presence of words. As the features aren't continuous variables, correlation analysis was less relevant and thus not performed.
- **Imbalanced Dataset:**
    - **Class Representation:** The dataset is completely balanced with an equal number of instances in class '0' (negative) and '1' or initially '4'(positive) each having 800000 instances.



3. Dataset Pre-processing:

In our sentiment analysis task, we conducted some of the standard text pre-processing methods mentioned below using the NLTK, Pandas and Regular Expressions (re)

- **Column Handling:**

  Initially, the dataset contained columns such as 'id', 'date', 'query', and 'user', which were deemed unnecessary for sentiment analysis and hence were removed from the dataset.

- **Label Transformation:**

  The sentiment labels originally ranged between 0 to 4. To simplify the analysis and create a binary classification task, the labels were transformed. The label '4' indicating a positive sentiment was replaced with '1', thereby converting the sentiment labels to binary—0 for negative sentiment and 1 for positive sentiment.

- **Text Cleaning:**

  Twitter-specific elements like handles (@user) were removed from the tweet text using regular expressions to eliminate user

mentions, focusing solely on the content.

	label	id	tweet	clean_tweet
0	0	1467810672	is upset that he can't update his Facebook by ...	is upset that he can't update his Facebook by ...
1	0	1467810917	@Kenichan I dived many times for the ball. Man...	I dived many times for the ball. Managed to s...
2	0	1467811184	my whole body feels itchy and like its on fire	my whole body feels itchy and like its on fire
3	0	1467811193	@nationwideclass no, it's not behaving at all....	no, it's not behaving at all. i'm mad. why am...
4	0	1467811372	@Kwesidei not the whole crew	not the whole crew

- **Text Normalization:**

  Special characters, numbers, and punctuation were removed to streamline the text data, ensuring uniformity in analysis. Short words were also removed to focus on more contextually meaningful content.

  Removal of special characters:

	label	id	tweet	clean_tweet
0	0	1467810672	is upset that he can't update his Facebook by ...	is upset that he can t update his Facebook by ...
1	0	1467810917	@Kenichan I dived many times for the ball. Man...	I dived many times for the ball Managed to s...
2	0	1467811184	my whole body feels itchy and like its on fire	my whole body feels itchy and like its on fire
3	0	1467811193	@nationwideclass no, it's not behaving at all....	no it s not behaving at all i m mad why am...
4	0	1467811372	@Kwesidei not the whole crew	not the whole crew

  Removal of short words:

	label	id	tweet	clean_tweet
0	0	1467810672	is upset that he can't update his Facebook by ...	upset that update Facebook texting might resul...
1	0	1467810917	@Kenichan I dived many times for the ball. Man...	dived many times ball Managed save rest bounds
2	0	1467811184	my whole body feels itchy and like its on fire	whole body feels itchy like fire
3	0	1467811193	@nationwideclass no, it's not behaving at all....	behaving here because over there
4	0	1467811372	@Kwesidei not the whole crew	whole crew

- **Tokenization:**

The tweet texts were tokenized, where individual tweets were split into a list of words (tokens) using whitespace as the separator. This step aimed to break down the tweets into smaller units for further analysis.

```
0    [upset, that, update, Facebook, texting, might...
1    [dived, many, times, ball, Managed, save, rest...
2              [whole, body, feels, itchy, like, fire]
3               [behaving, here, because, over, there]
4                                             [whole, crew]
Name: clean_tweet, dtype: object
```

- **Stemming:**

  The Porter Stemmer algorithm was employed to reduce words to their root form, aiming to normalize variations of words and reduce the overall vocabulary size.

```
0    [upset, that, updat, facebook, text, might, re...
1    [dive, mani, time, ball, manag, save, rest, bo...
2               [whole, bodi, feel, itchi, like, fire]
3                   [behav, here, becaus, over, there]
4                                          [whole, crew]
Name: clean_tweet, dtype: object
```

- **Feature Extraction:**

  The cleaned and pre-processed text data was transformed using the Bag-of-Words (BoW) technique, converting text into numerical vectors, facilitating the application of machine learning models. We applied several filtering steps: excluding terms appearing in over 90% of the tweets, filtering out very rare words appearing in fewer than two tweets, and limiting the features to the top 1000 most frequent words in the corpus. Additionally, we eliminated common English stop words ('the', 'and', 'is', etc.) that lack significant sentiment information. This process resulted in a

matrix with dimensions (1600000, 1000), symbolizing 1,600,999 tweets and 1000 unique words as features for subsequent analysis.

4. Feature Scaling

- Feature scaling is not required for the Bag-of-Words representation that we used, because the values in the feature vectors are binary indicators. Each feature represents the presence of a specific word in the document, and these counts are already on the same scale.

5. Dataset Splitting

- **Random Split**: Data was split randomly into 70% training and 30% testing sets.

6. Model Training & Testing

We employed two classification models to fit our data, logistic regression which is a discriminative classifier and Naive Bayes which is a generative classifier.

- **Logistic Regression**
  - Training: During the training phase, the logistic regression model was fitted on the training data derived from the Bag-of-Words (BoW) representation of tweet text. This process involved adjusting the model's parameters to minimize the difference between predicted and actual sentiments.

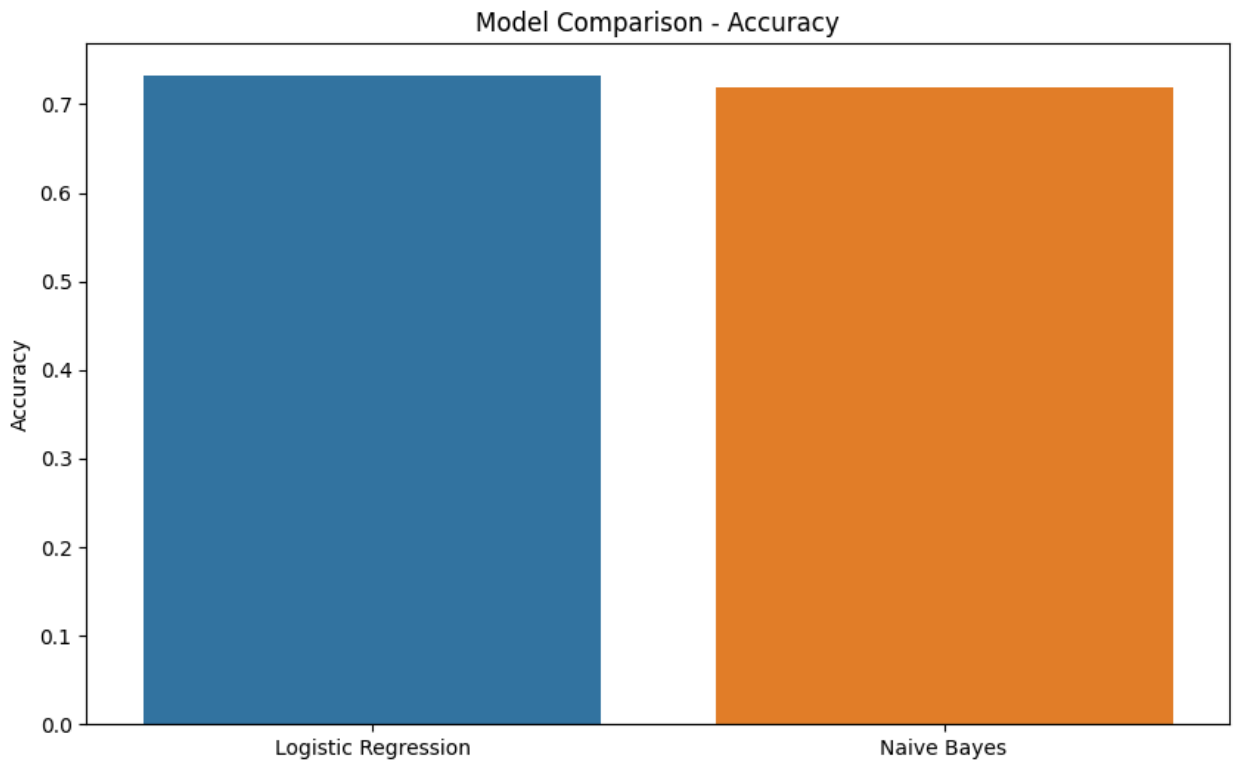
- Testing: The trained logistic regression model was then evaluated using the separate test dataset. Predictions were made on this unseen data to assess the model's performance in classifying sentiments, measuring metrics like accuracy and F1 score.

- **Naive Bayes**
  - Training: In the training phase for Naive Bayes, the model learned the statistical probabilities of different words in the dataset. This probabilistic approach assumes independence among features and calculates the likelihood of a sentiment given the presence of certain words.
  - Testing: Following training, the Naive Bayes classifier was tested on the test dataset to estimate its performance. It predicted sentiments based on the learned probabilities and assessed its accuracy and other relevant metrics.

7. Model Selection/Comparison Analysis

- **Comparison of Models**
  - Accuracy

    The bar chart below illustrates the comparison of accuracy between the Logistic Regression and Naive Bayes models. Both models have been trained and tested on the sentiment analysis dataset, and their performances are evaluated based on their accuracy in predicting sentiment labels.



The accuracy comparison suggests that the Logistic regression model having accuracy of 73.185% outperforms the Naive Bayes Model having an accuracy of 71.83%, in terms of overall prediction accuracy by a small percentage.

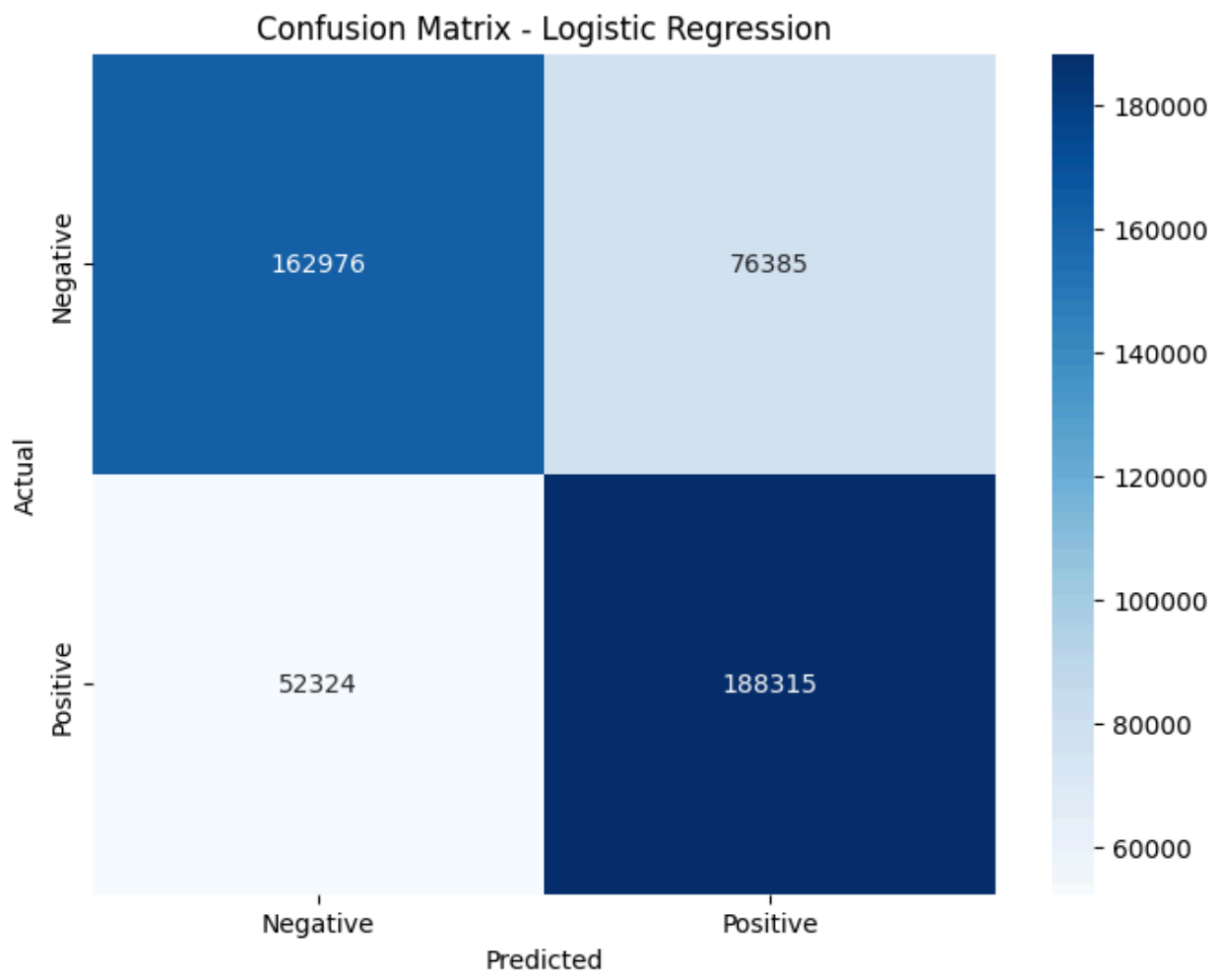
- **Performance Metrics Comparison**
  - Precision, Recall, F1 Score:

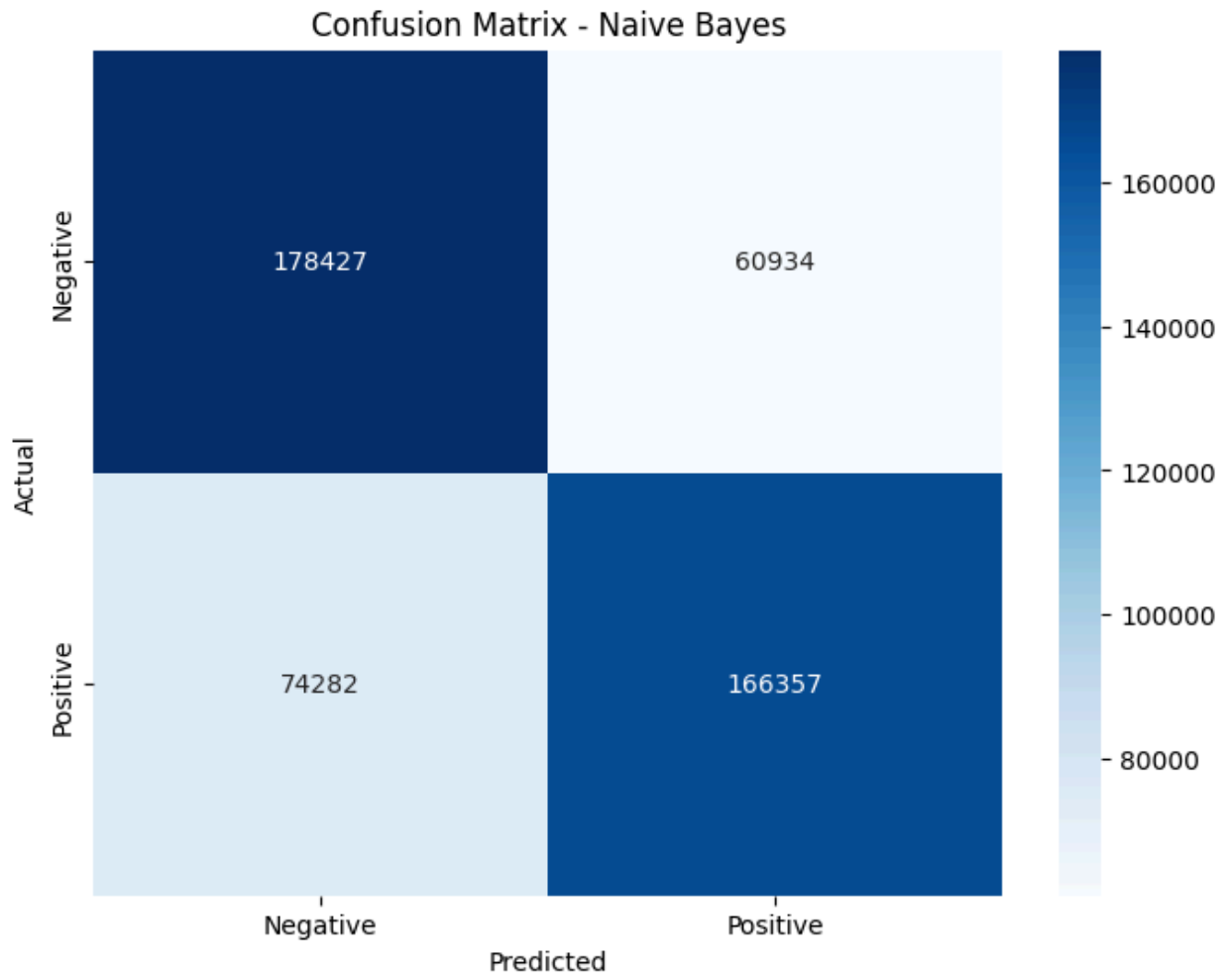
The table below provides a detailed comparison of precision, recall, and F1 score for both models. These metrics offer insights into the models' ability to correctly classify positive and negative sentiments.Even though the Naive Bayes Model has a higher precision the overall F1 Score suggests that the Logistic Regression model is better.

Model	Precision	Recall	F1 Score
Logistic Regression	0.711428	0.782562	0.745302
Naive Bayes	0.731912	0.691314	0.711034

- **Confusion Matrix**
  - confusion matrix for logistic regression and naive bayes:

    The confusion matrix for the Logistic Regression model and Naive Bayes model provides a detailed breakdown of its predictions. It illustrates the number of instances classified as true positives, true negatives, false positives, and false negatives.


Confusion Matrix - Logistic Regression
Negative
Positive
Actual
162976
76385
52324
188315
Negative
Positive
Predicted
180000
160000
140000
120000
100000
80000
60000



8. Conclusion

The sentiment analysis project successfully explored and classified sentiments in Twitter data, employing techniques such as data exploration, preprocessing, and machine learning models. Notably, the logistic regression model demonstrated promising performance in sentiment classification. The project utilized the Bag-of-Words (BoW) representation for feature extraction and compared the performance of logistic regression and Naive Bayes models.

In summary, the comprehensive analysis provided valuable insights into sentiments expressed in user-generated text related to artificial intelligence

(AI). The combination of exploratory data analysis, preprocessing techniques, and machine learning models contributed to a deeper understanding of public sentiments towards AI technologies, applications, and developments. This sentiment analysis project serves as a robust foundation for further research and decision-making in the AI domain.