SVKM'S

NMIMS

Deemed to be UNIVERSITY+

Applied Artificial Intelligence

A report on **Sentiment Analysis Report: Malaysia**

School Of Technology & Management Engineering

Department of CSE – DS

Jadcherla , Telangana , 509301

Submitted By

Gudyagopu Pragnya Reddy

Roll no. : L055

Sentiment Analysis Report: Malaysia

Introduction :

This report presents a sentiment analysis of text data extracted from Wikipedia's Malaysia page. Using natural language processing (NLP) techniques and machine learning algorithms, the analysis examines sentiment patterns and identifies key topics in the content. The study employs sentiment classification, word frequency analysis, and predictive modeling to understand how Malaysia is portrayed in this widely-accessed information source. Wikipedia shapes public understanding of nations and cultures globally. By analyzing Malaysia's page, this study reveals insights into the digital representation of this diverse Southeast Asian nation with its unique constitutional monarchy and multicultural society. The research follows a structured approach: web scraping for data collection, text preprocessing, sentiment analysis at the sentence level, and machine learning classification, providing a comprehensive picture of Malaysia's representation in encyclopedic content.

Objectives :

1. To determine the overall sentiment distribution (positive, negative, and neutral) in the Wikipedia content related to Malaysia
2. To identify and analyze key topics, themes, and frequently occurring words within the text
3. To evaluate the performance of machine learning models in classifying sentiment in Malaysia-related content
4. To uncover patterns in how Malaysia's governance, culture, geography, and other aspects are portrayed in encyclopedic content
5. To provide data-driven insights that could inform understanding of public information representation of Malaysia

This sentiment analysis study offers several significant contributions:

1. Information Representation: The findings reveal how encyclopedic content portrays a nation, providing insights into potential biases or emphasis in publicly available information about Malaysia.
2. Methodological Value: The project demonstrates the application of NLP and machine learning techniques to analyze large volumes of text data, showcasing both the capabilities and limitations of these approaches.
3. Benchmark Creation: The results establish a benchmark for sentiment distribution in encyclopedic content about Malaysia, which can be compared with other countries or with future analyses of the same content as it evolves.

Data Collection :

The data collection process for this sentiment analysis project followed a systematic approach to gather comprehensive textual information about Malaysia. The process began with web scraping the Wikipedia page on Malaysia (https://en.wikipedia.org/wiki/Malaysia) using Python's BeautifulSoup library. The initial scraping retrieved approximately 61,389 characters of raw text content from the main body of the article.

After retrieval, the text underwent cleaning procedures to prepare it for analysis. This included removing citations (such as [1], [2]), eliminating special characters and digits, and normalizing whitespace. The cleaning process resulted in 58,091 characters of processed text, representing a 5.4% reduction from the original content while preserving the semantic information.

The cleaned text was then tokenized into 470 individual sentences using NLTK's sentence tokenizer. These sentences formed the primary units of analysis for sentiment classification. Word tokenization was subsequently applied, yielding 9,619 total words, which were further reduced to 5,188 meaningful words after removing stopwords and non-alphabetic tokens. This collection methodology ensured that the dataset contained rich, domain-specific information about Malaysia's geography, government, culture, and history from a reliable encyclopedia source, providing sufficient material for meaningful sentiment and topic analysis.

Data Preprocessing :

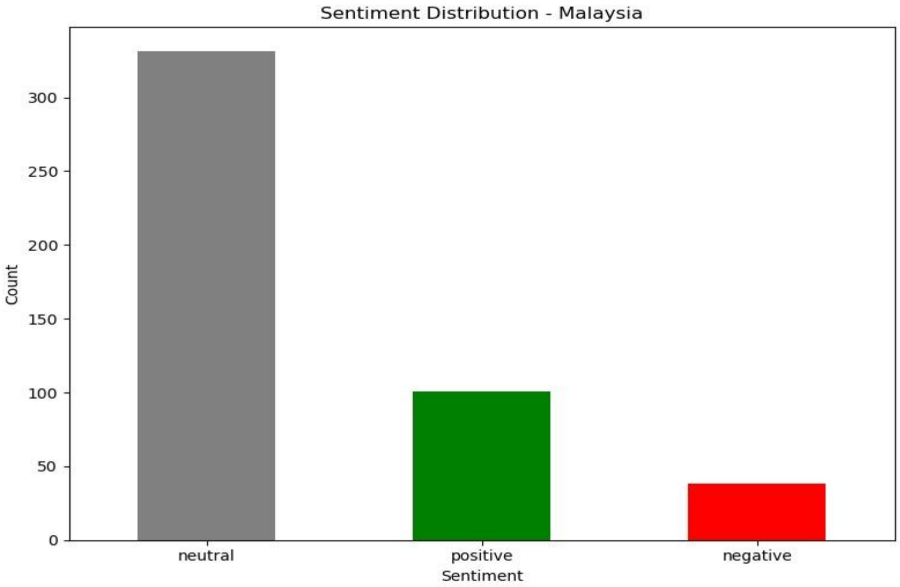
The raw textual data extracted from Wikipedia underwent several preprocessing steps to transform it into a suitable format for analysis. First, regular expressions were applied to remove citations (text in square brackets like [1], [2]) that are common in Wikipedia articles but don't contribute to semantic meaning. Special characters, digits, and non-alphabetic symbols were also eliminated to reduce noise, while maintaining periods for sentence boundaries. Multiple whitespaces were normalized to single spaces to create a cleaner text structure. This initial cleaning reduced the character count from 61,389 to 58,091 while preserving the essential content. The cleaned text was then tokenized into 470 distinct sentences using NLTK's sentence tokenizer, creating the fundamental units for sentiment analysis.

For word-level analysis, the text was converted to lowercase and tokenized into individual words, yielding 9,619 tokens. Stopwords (common words like "the," "and," "is") were removed using NLTK's English stopword list, reducing the word count to 5,188 meaningful content words that better represent the topical substance of the text. TextBlob was employed to calculate sentiment polarity for each sentence, with values ranging from -1 (highly negative) to +1 (highly positive). These polarity scores were then categorized into three sentiment classes: positive (>0.1), negative (<-0.1), and neutral (-0.1 to 0.1). This classification resulted in 331 neutral, 101 positive, and 38 negative sentences. For machine learning purposes, the text was vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) with a maximum of 5,000 features and minimum document frequency of 2, resulting in a feature matrix of shape (139, 318) for non-neutral sentences.

Sentiment Analysis of Malaysia’s Wikipedia Page:

Sentiment Distribution : The sentiment analysis of the Wikipedia content on Malaysia reveals a strong tendency toward neutral language:

- Neutral: 331 sentences (70.4%)
- Positive: 101 sentences (21.5%)
- Negative: 38 sentences (8.1%)

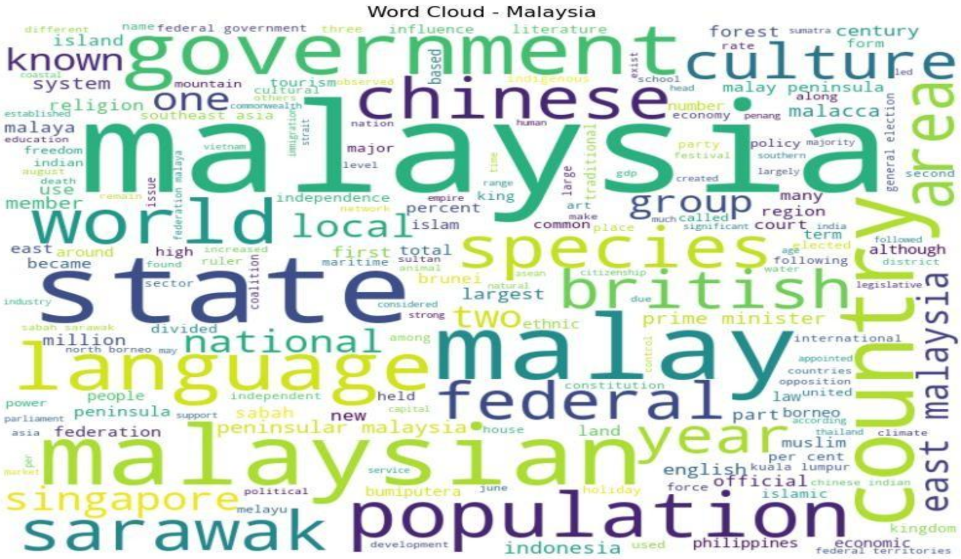


This distribution suggests that the content adheres closely to Wikipedia's editorial standards, which emphasize objectivity and neutrality. The relatively higher number of positive sentences compared to negative ones may indicate a generally favorable or balanced portrayal of Malaysia's features, achievements, and socio-political environment.

Key Topics and Word Frequency : The word cloud highlights the most prominent terms related to Malaysia, with these terms appearing most frequently:

1. "Malaysia"

   (most dominant)
2. "government"
3. "state"
4. "federal"
5. "malay"
6. "malaysian"
7. "chinese"



These terms emphasize the core themes in the content, primarily related to governance, ethnicity, and geography.

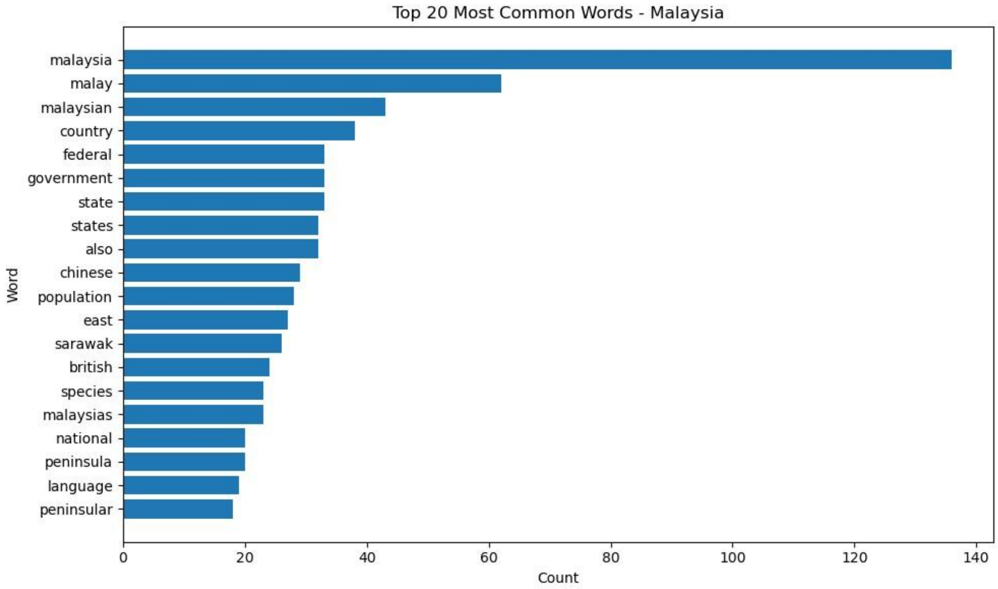
The horizontal bar chart quantitatively supports these observations:

- "Malaysia" appears approximately 140 times
- "Malay" appears around 60 times
- "Malaysian" appears about 40 times

These findings indicate a strong focus on:

- Governance and political structure (e.g., *government*, *federal*, *state*)
- Cultural and ethnic composition (e.g., *Malay*, *Chinese*, *Malaysian*)
- Geographical features (e.g., *peninsula*, *east*, *Sarawak*)

The prominence of these topics reflects Wikipedia's comprehensive coverage of Malaysia's political, cultural, and geographic landscape.



Machine Learning Classification Models

To evaluate the capability of machine learning in classifying sentiment, two models were implemented—Logistic Regression and Naive Bayes—on the dataset after excluding neutral sentences. The objective was to classify the remaining sentences as either *positive* or *negative*.

○ Logistic Regression Model

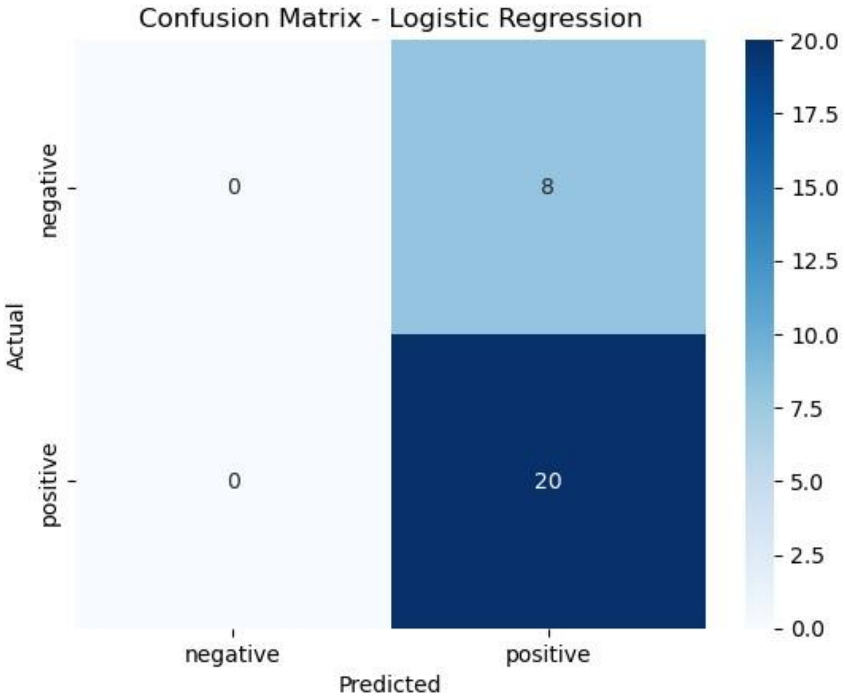
The performance of the Logistic Regression model is visualized and summarized below:

- Accuracy: 71%
- Precision (Positive class): 0.71
- Recall (Positive class): 1.00
- F1-score (Positive class): 0.83
- Negative class metrics: 0.00 across all metrics (no negative instances correctly classified)

The confusion matrix (Image 5) further illustrates the performance:

- Correctly predicted 20 positive instances
- Misclassified all 8 negative instances as positive

This indicates the model is biased toward the positive class, failing entirely to detect negative sentiment.



○ Naive Bayes Model

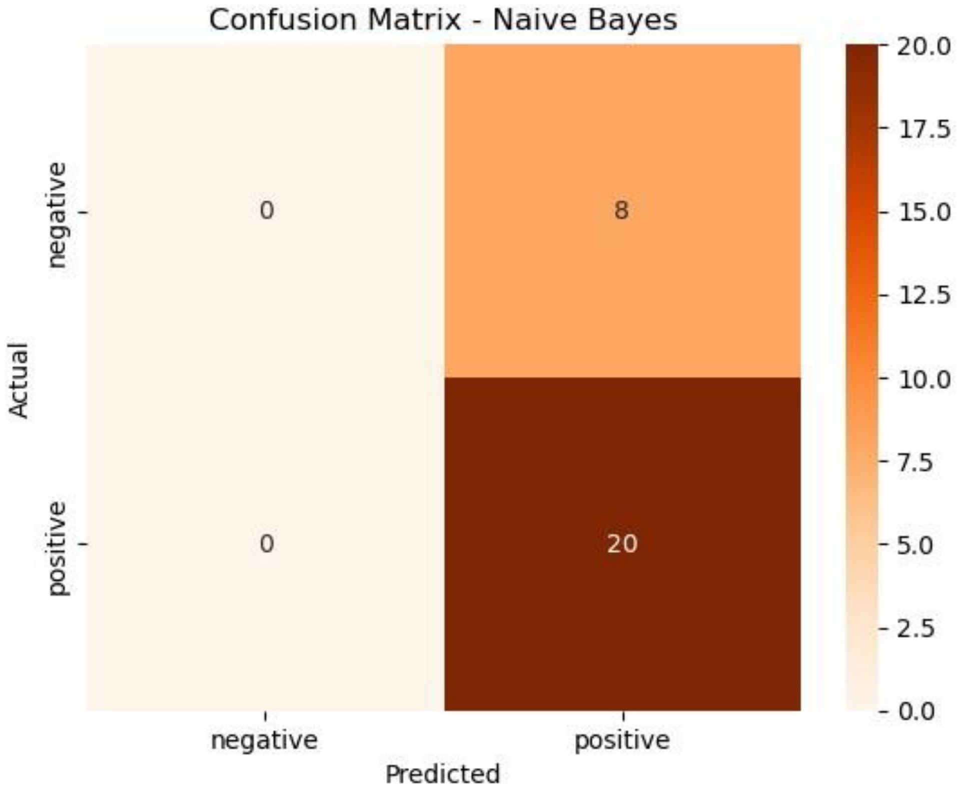
The Naive Bayes model (results shown in Image 6) yielded identical performance metrics as Logistic Regression:

- Accuracy: 71%
- Precision (Positive class): 0.71
- Recall (Positive class): 1.00
- F1-score (Positive class): 0.83
- Negative class metrics: 0.00

The confusion matrix (Image 7) confirms the same pattern:

- All positive instances were correctly identified
- All negative instances were misclassified as positive

These results suggest both models are ineffective at recognizing negative sentiment, likely due to class imbalance and limited negative training data. This highlights the need for data balancing techniques or advanced modeling approaches (e.g., ensemble methods or SMOTE) for improved classification of minority sentiment classes.



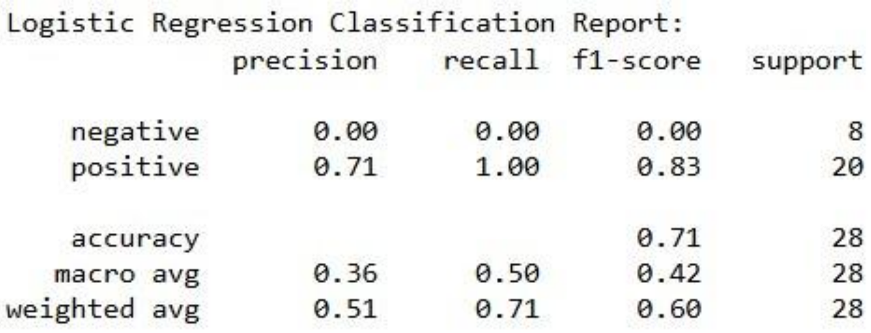
Analysis of Classification Results :

The outcomes of both the Logistic Regression and Naive Bayes models reveal several critical insights regarding the limitations and challenges in sentiment classification for this dataset:

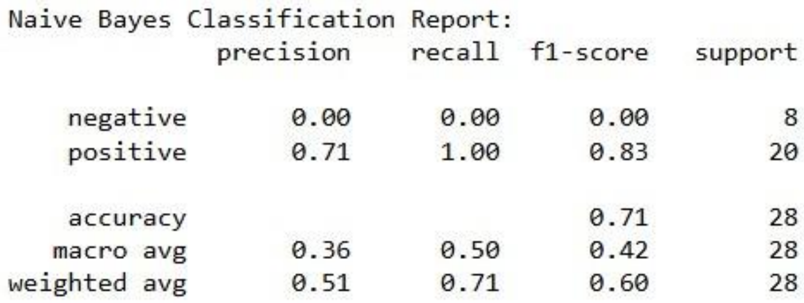
Class Imbalance Impact : The dataset contains a significantly higher number of positive instances (101) compared to negative instances (38). This class imbalance skews model learning, leading to an overemphasis on the majority class. As a result, the models achieve seemingly good overall accuracy but fail to perform well on the minority (negative) class.

Bias Toward Majority Class : Both models exhibit a strong bias toward the positive class. This is evident in the confusion matrices, where all negative instances were misclassified as positive. While this results in perfect recall for the positive class, it comes at the cost of zero recall and precision for the negative class, rendering the models ineffective for balanced sentiment detection.

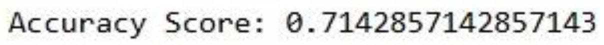
Limited Feature Discrimination : The use of TF-IDF features, while effective in general text classification tasks, may lack the granularity needed to distinguish subtle linguistic cues between positive and negative sentiments. This limitation restricts the models' ability to accurately capture sentiment polarity, especially when the vocabulary used in both classes overlaps significantly.



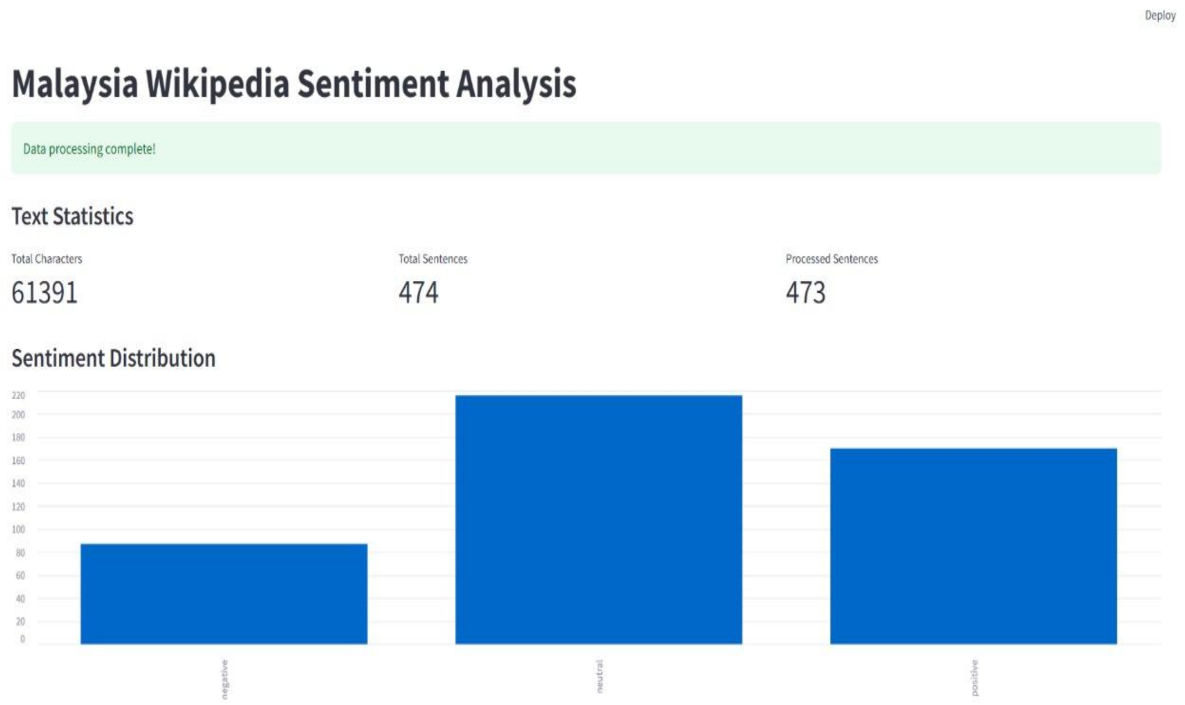
```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00         8
    positive       0.71      1.00      0.83        20

    accuracy                           0.71        28
   macro avg       0.36      0.50      0.42        28
weighted avg       0.51      0.71      0.60        28
```



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

    negative       0.00      0.00      0.00         8
    positive       0.71      1.00      0.83        20

    accuracy                           0.71        28
   macro avg       0.36      0.50      0.42        28
weighted avg       0.51      0.71      0.60        28
```



```
Accuracy Score: 0.7142857142857143
```



Malaysia Wikipedia Sentiment Analysis

Data processing complete!

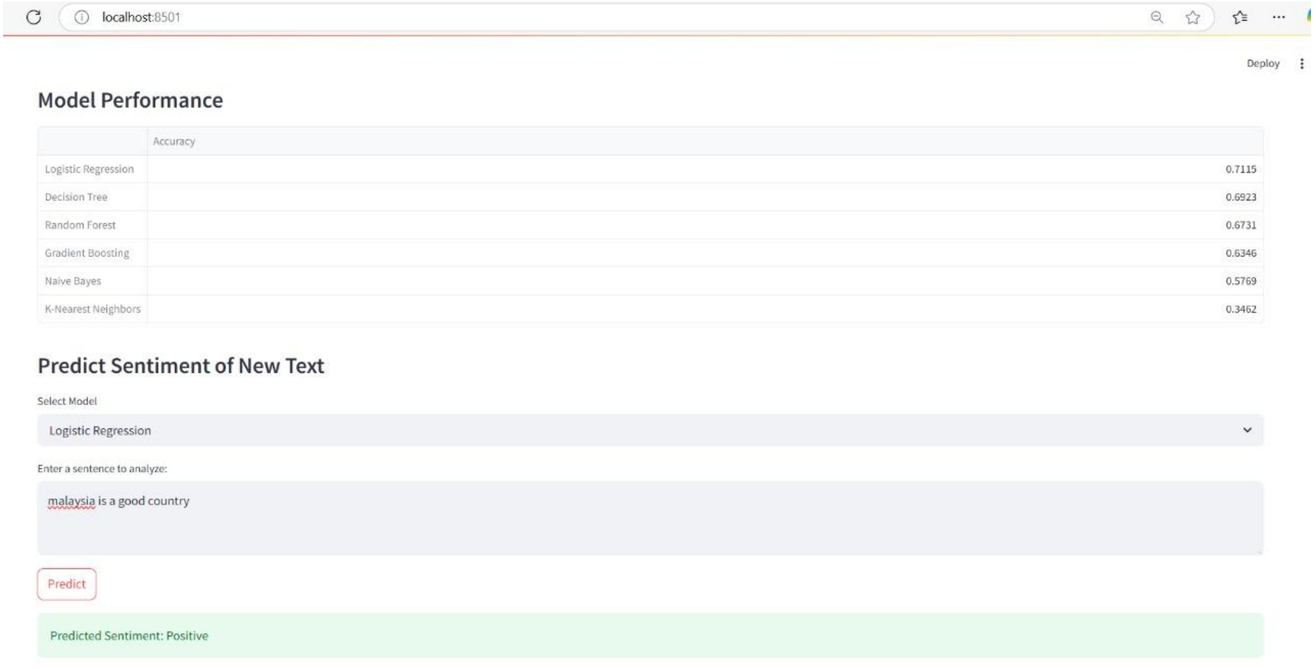
Text Statistics

Total Characters
61391

Total Sentences
474

Processed Sentences
473

Sentiment Distribution



Model Performance

	Accuracy
Logistic Regression	0.7115
Decision Tree	0.6923
Random Forest	0.6731
Gradient Boosting	0.6346
Naive Bayes	0.5769
K-Nearest Neighbors	0.3462

Predict Sentiment of New Text

Select Model

Logistic Regression

Enter a sentence to analyze:

malaysia is a good country

Predicted Sentiment: Positive

Conclusion :

The sentiment analysis of Wikipedia content about Malaysia indicates that the text is predominantly neutral, with a higher occurrence of positive sentiments compared to negative ones. This aligns with Wikipedia’s objective, fact-based editorial style. The analysis of key terms reveals a focus on governance, cultural diversity, and geographical aspects, suggesting that the coverage of Malaysia is both balanced and comprehensive.

The machine learning models—Logistic Regression and Naive Bayes—demonstrated moderate overall accuracy but failed to accurately classify negative sentiment due to class imbalance and limited feature discrimination. These results underscore the importance of addressing data imbalance and enhancing feature representation in sentiment classification tasks to achieve more robust and reliable performance across all sentiment classes.