Sentiment Analysis of Tweets Using Machine Learning

PRESENTED BY:
Elisha Damor

AGENDA

Stage 1
- Introduction & Importance of Sentiment Analysis

Stage 2
- Overview of Sentiment140 Dataset

Stage 3
- Data Preprocessing Steps
- Introduction to Machine Learning Models Used

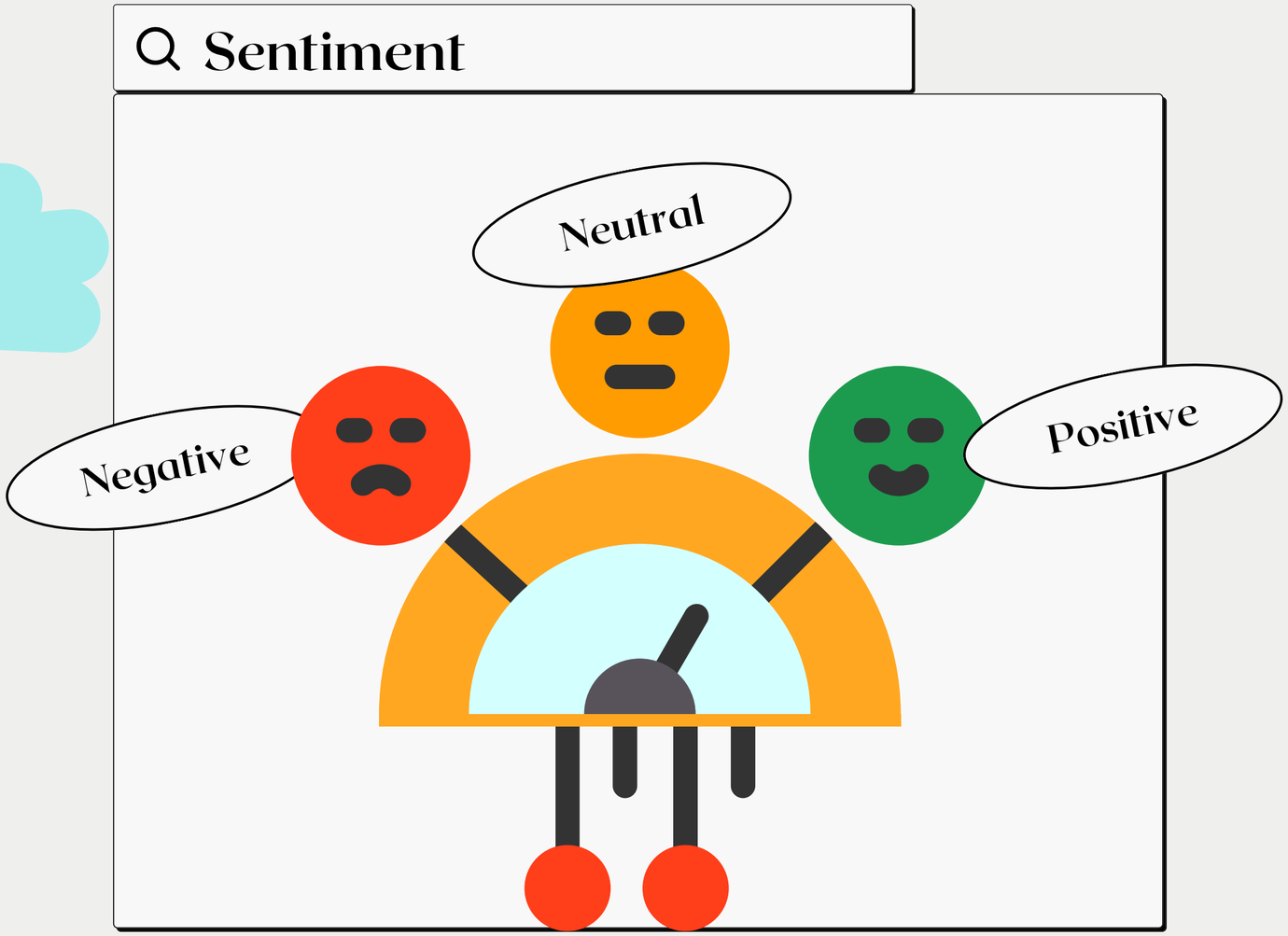
Stage 4
- Comparison of Model Performance
- Best Model

Stage 5
- Challenges Faced
- Future Work
- Conclusions

Stage 6
- Applications of Sentiment Analysis

Introduction to Sentiment Analysis/opinion mining

Sentiment analysis involves classifying text into positive, negative, or neutral categories to understand public sentiment.



Importance of Sentiment Analysis

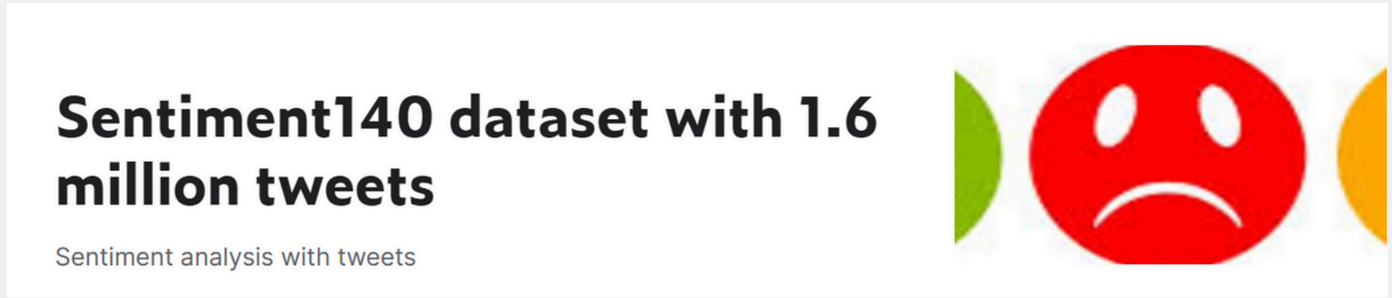
Why sentiment analysis matters? Sentiment analysis is crucial for market analysis, political campaigns, and public relations to gauge public opinion and improve strategies.

Marketing

Customer Service

Political Campaigns

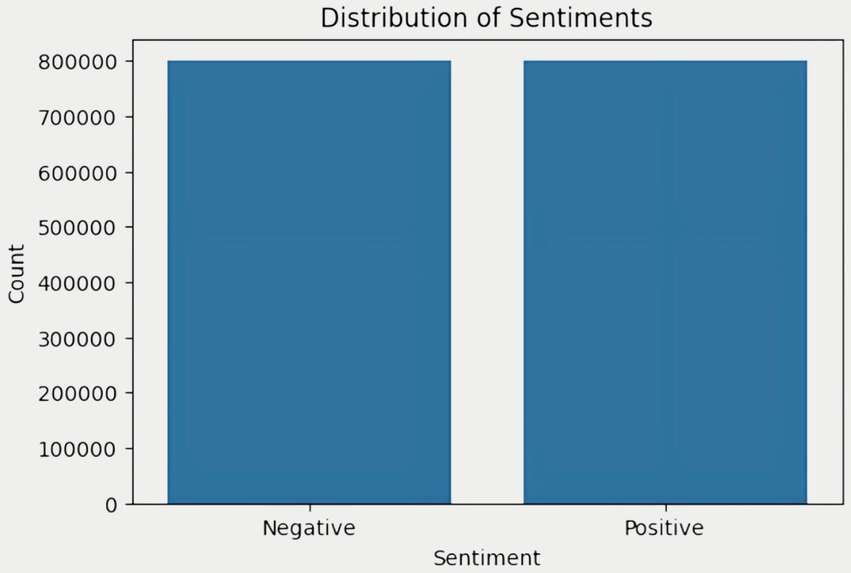
Overview of Sentiment140 Dataset



- Created by researchers at Stanford University
- Contains 1.6 million tweets labeled as positive or negative
- Captures real-time public sentiments from x

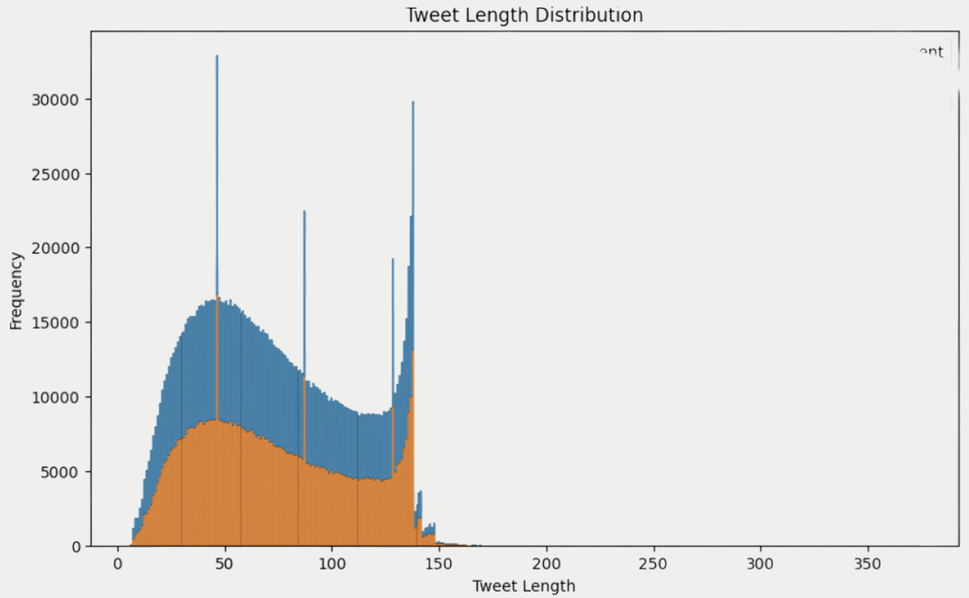
Exploratory Data Analysis (EDA)

Sentiment Distribution



Shows an equal number of positive and negative tweets, indicating a balanced dataset.

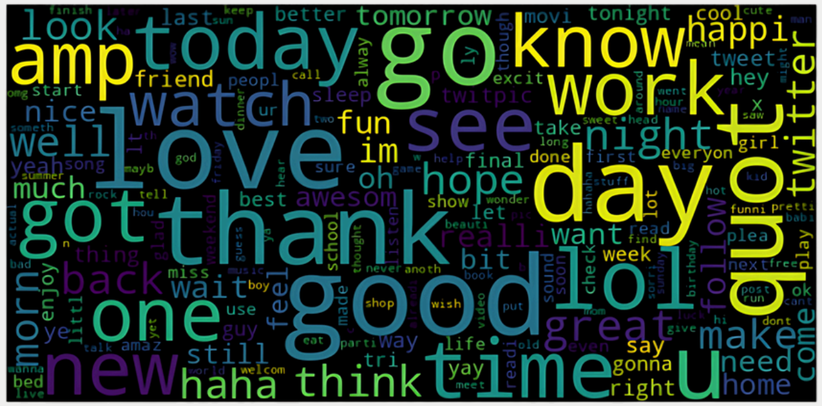
Tweet Length Distribution



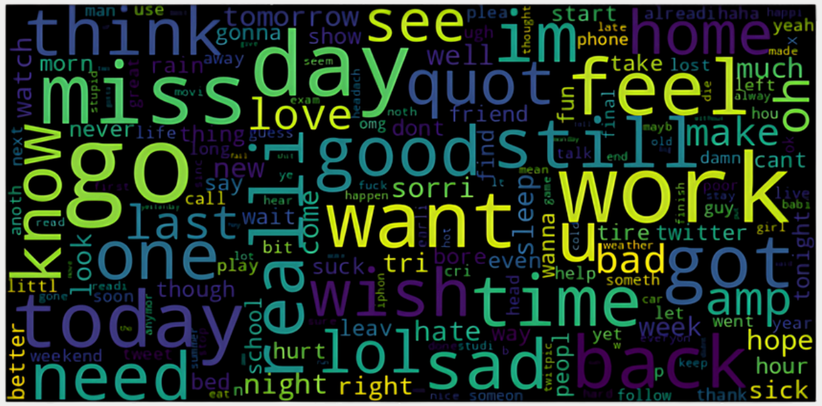
Most tweets are around the 140-character limit, which is Twitter's character limit

Exploratory Data Analysis ((EDA)

Word Clouds



Positive Tweets Word Cloud



Negative Tweets Word Cloud

Data Preprocessing

Text Cleaning

Removed URLs, usernames, special characters, and numbers.

Stop Words Removal

Removed stop words like 'and', 'the', 'is', which don't add much value to sentiment analysis.

Data Preprocessing

Encoding Labels

```
# Convert string labels to integers
encoder = LabelEncoder()
Y_test_encoded = encoder.fit_transform(Y_test)
Y_pred_encoded = encoder.transform(Y_pred)
```

Used LabelEncoder to convert categorical labels to numerical values.

Stemming

'JUMPS', 'JUMPED',

'JUMP'

Utilised Porter Stemming algorithm to reduce words to their root forms. For example, 'running' becomes 'run'. This helps in standardizing different forms of the same word.

Tokenization

Breaking text into tokens.

Feature Engineering

```
# Vectorization
vectorizer = TfidfVectorizer(max_features=5000)
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

Model Selection & Training

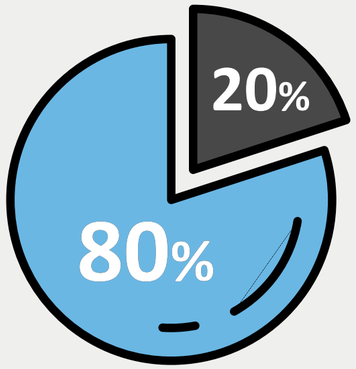
Logistic Regression

Decision Tree

LightGBM

Naive Bayes.

Training Method:



Standard 80/20 train–test split.

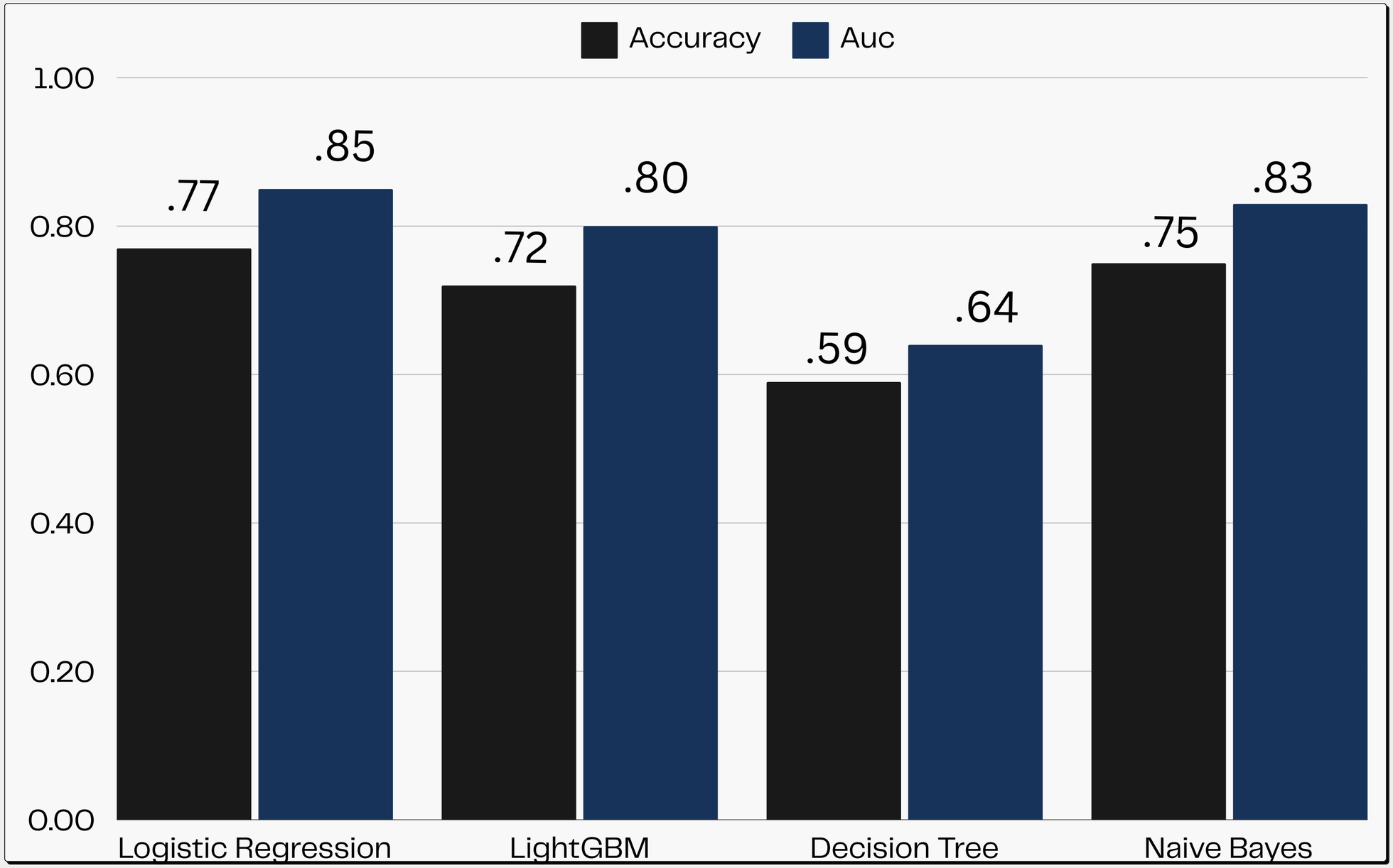
Cross-Validation / Model Evaluation

5 fold – Ensures robustness and prevents overfitting.

ACCURACY

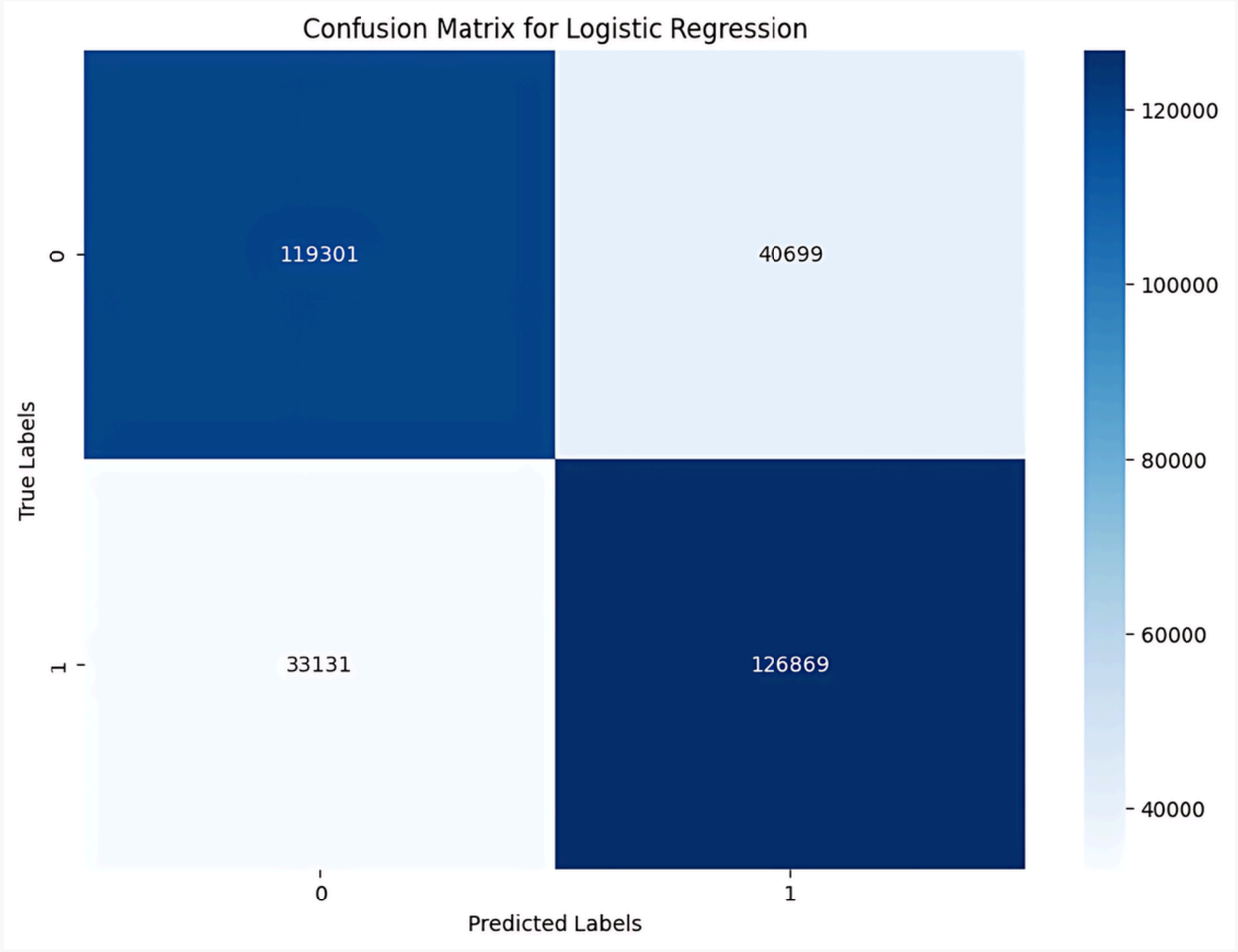
AUC

Competitive Analysis



Best Model

Logistic Regression



Hyperparameter Tuning

```
# Initialize the GridSearchCV
grid_search = GridSearchCV(estimator=lr_model,
param_grid=param_grid, cv=5, scoring='accuracy')

Best Parameters: {'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}
Logistic Regression Test Accuracy: 0.76863125
Logistic Regression Test AUC: 0.7686312499999999
```

Result

The Logistic Regression model with the identified best parameters shows robust performance both in cross-validation and on the test set.

Sentiment Analysis

```
# Example usage
sample_text = "I love using my new phone; it's awesome!"
print("Prediction:", preprocess_and_predict(sample_text))
```

```
Prediction: Positive
```

Chalenges

- Computational Overhead
- Limitation of computational power -

Libraries used-

NLTK,
Pandas, and Scikit-learn

Future Work

- Advanced Models: Exploring more advanced models such as transformer-based models (e.g., BERT) could provide even better results, albeit with higher computational costs.

- Unsupervised Learning: Integrating unsupervised learning techniques to capture more nuanced sentiments could enhance the robustness of sentiment analysis.

- Cross-Platform Analysis: Extending the analysis to other social media platforms could provide a more comprehensive understanding of public sentiment.

Real-world applications

Social Media Monitoring

- Use in Social Media
- Training on Pre-labeled Data
- Classifying Comments
- Business Strategy Adjustment

Real-world applications

Enhances


Thank You