Social Media Sentiment Analysis: Capturing Relationships Between Words and Sentiments

Amanda Altamirano, Data Analytics ’25; Abhik Raj Shrestha, Data Analytics/ International Business ‘25

INTRODUCTION

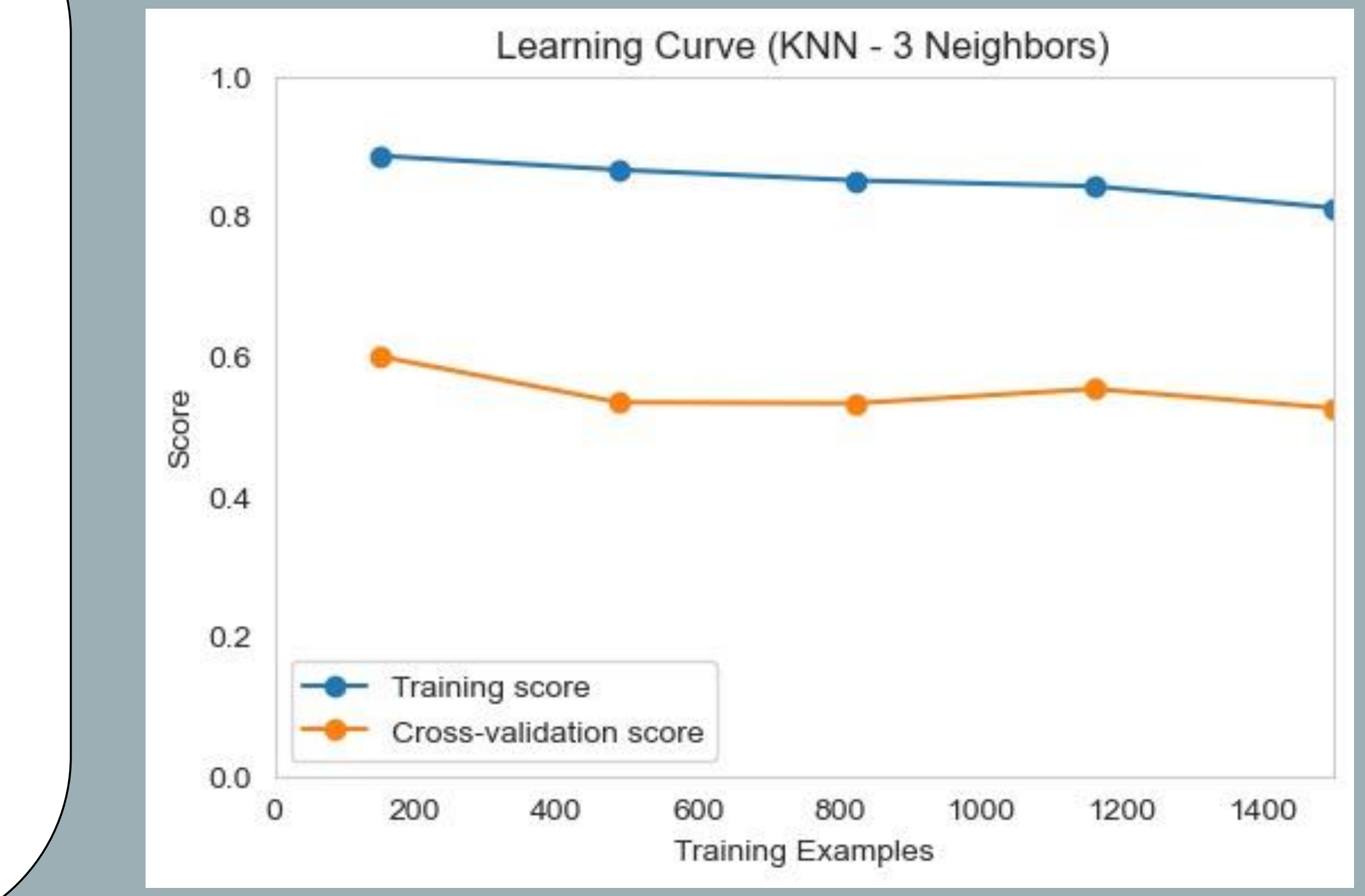
With the rapid rise of social media as a primary mode of communication, understanding public sentiment through online text has become increasingly important. This project focuses on analyzing the relationship between language and sentiment in social media posts using a labeled text dataset. We followed a three-step process: data preprocessing, visualization, and model testing. After cleaning and normalizing the data, we explored sentiment trends and word patterns through visualizations such as boxplots, bar charts, line plots, and scatter plots. We then tested and compared several machine learning models to classify sentiment, aiming to identify the most effective approach for capturing how emotional tone is reflected in word usage.

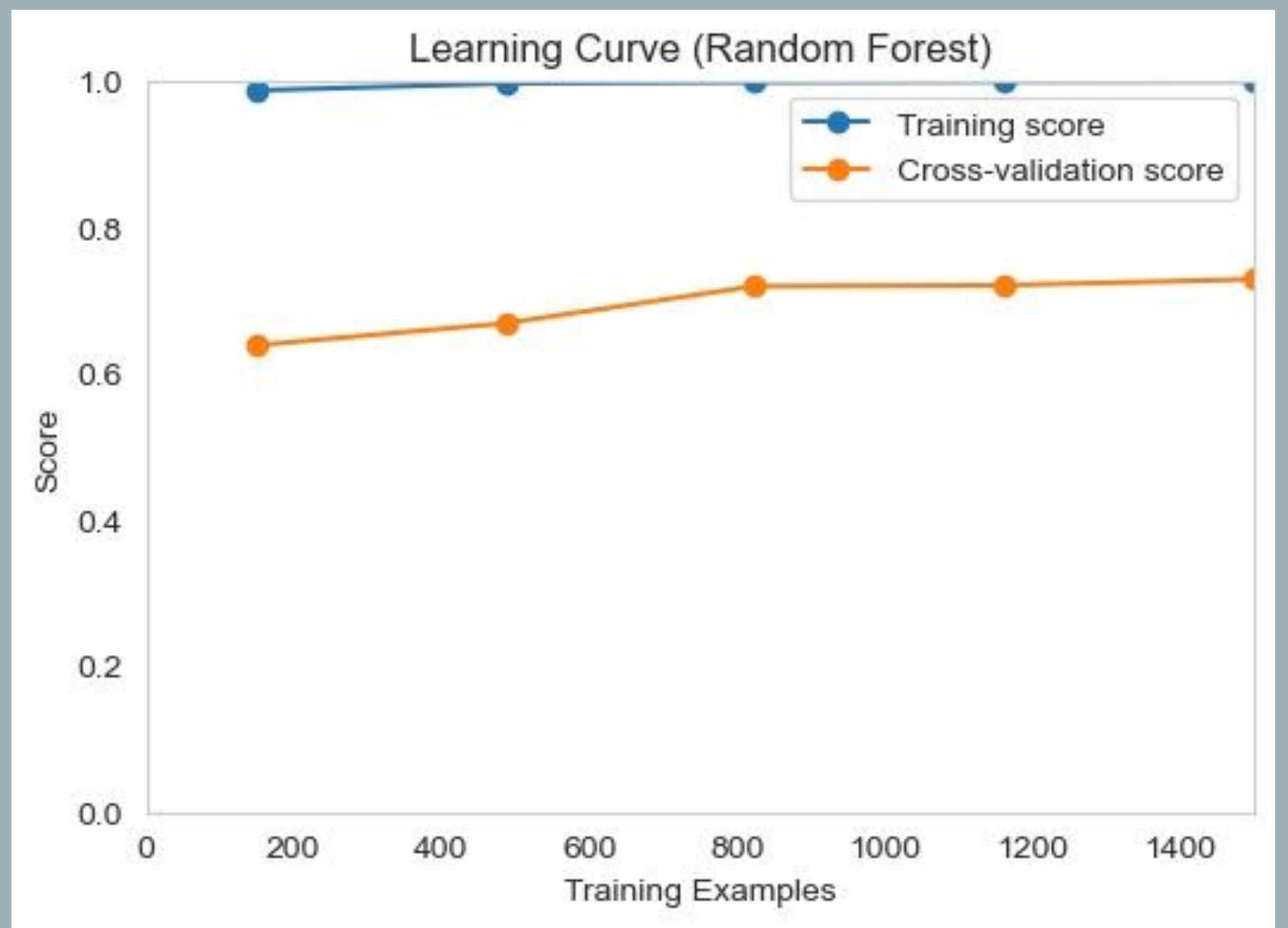
OBJECTIVES

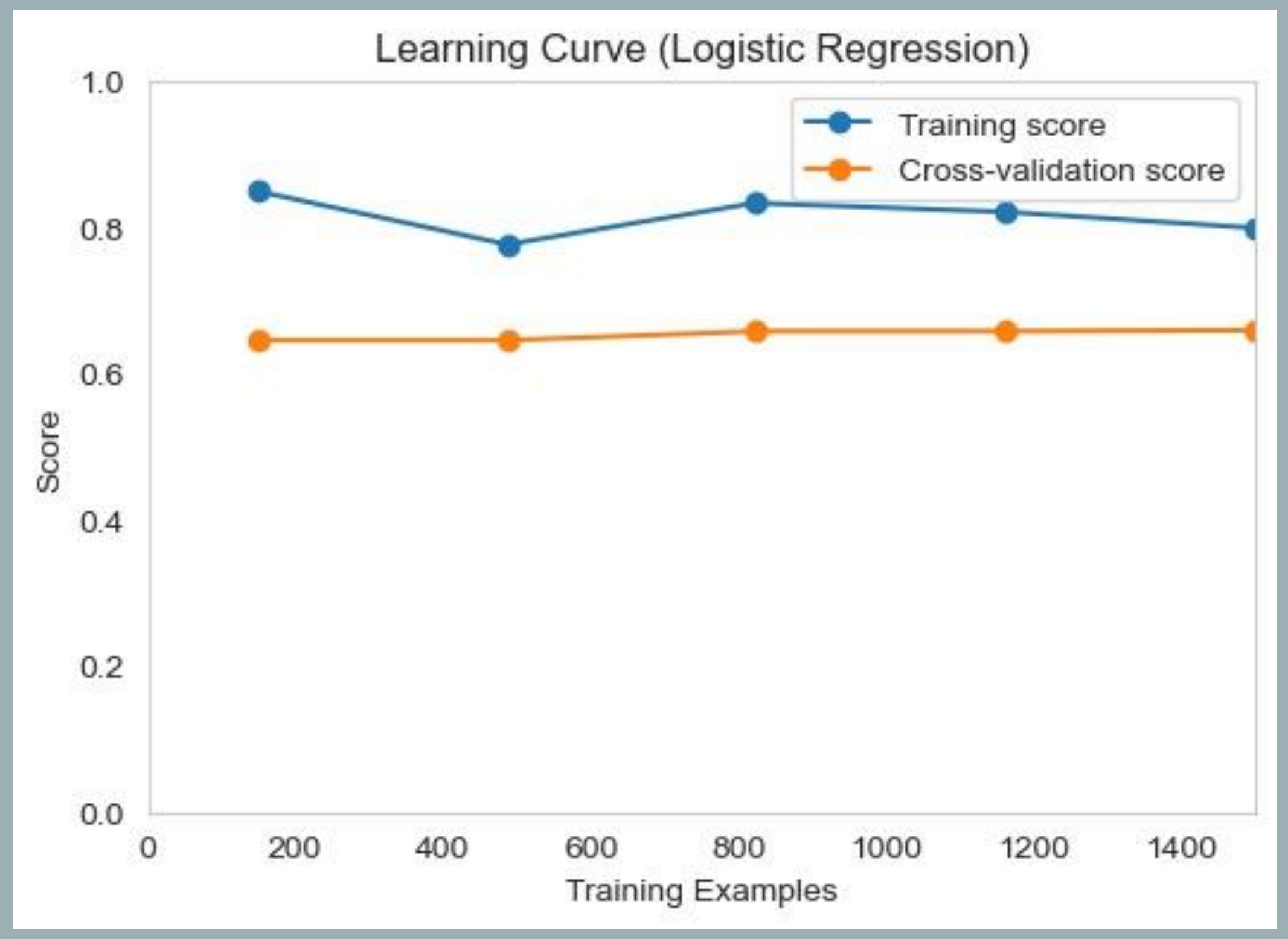
- To investigate specific word usage patterns correlates with emotional sentiment in social media posts.
- To evaluate and compare the performance of various machine learning models, including K-Nearest Neighbors (KNN), Logistic Regression, Naïve Bayes, and Random Forest, in accurately classifying sentiments.
- To demonstrate the practical value of sentiment analysis for businesses in monitoring customer feedback, brand perception, and market trends.

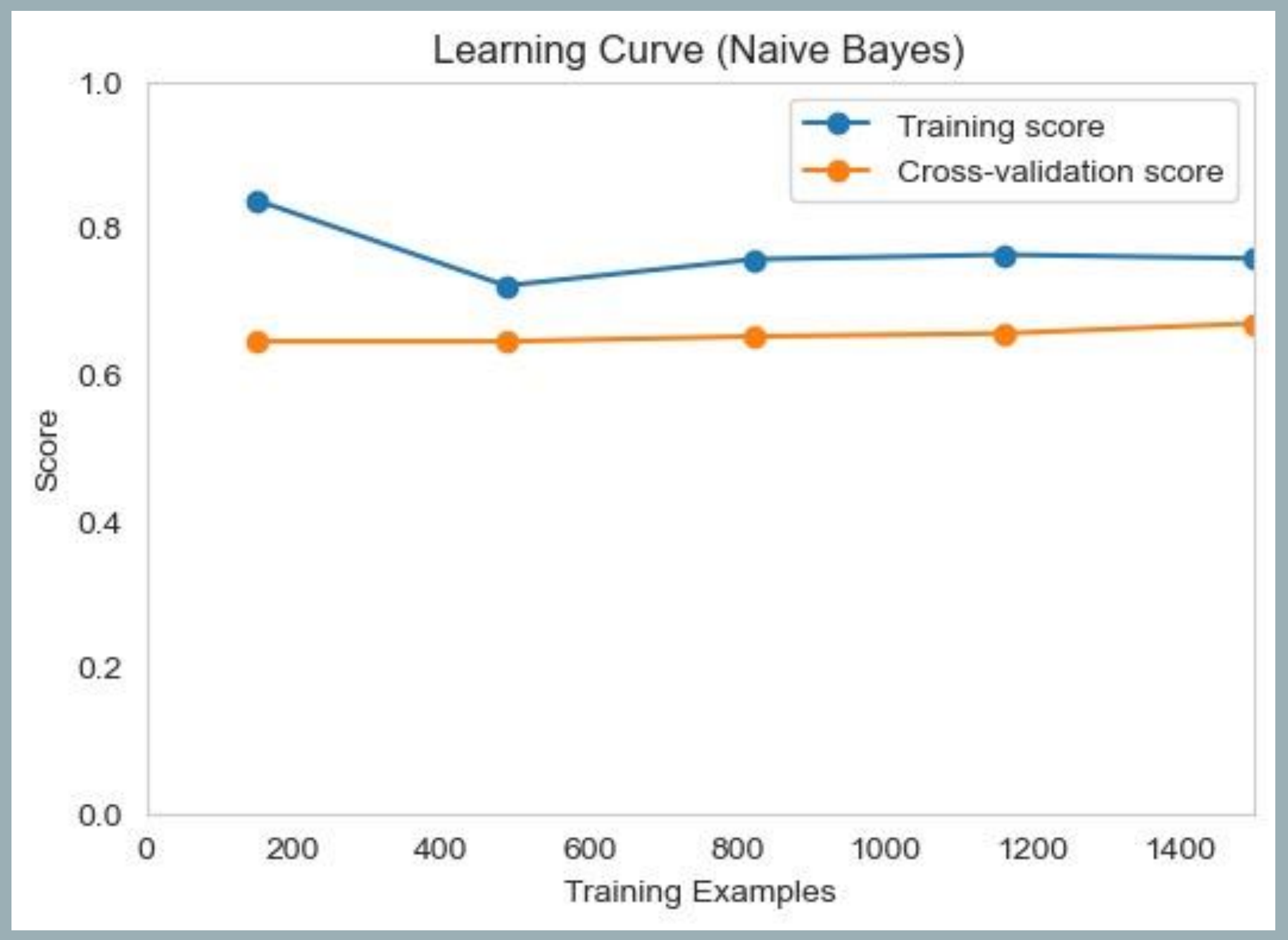
DATA AND METHODOLOGY

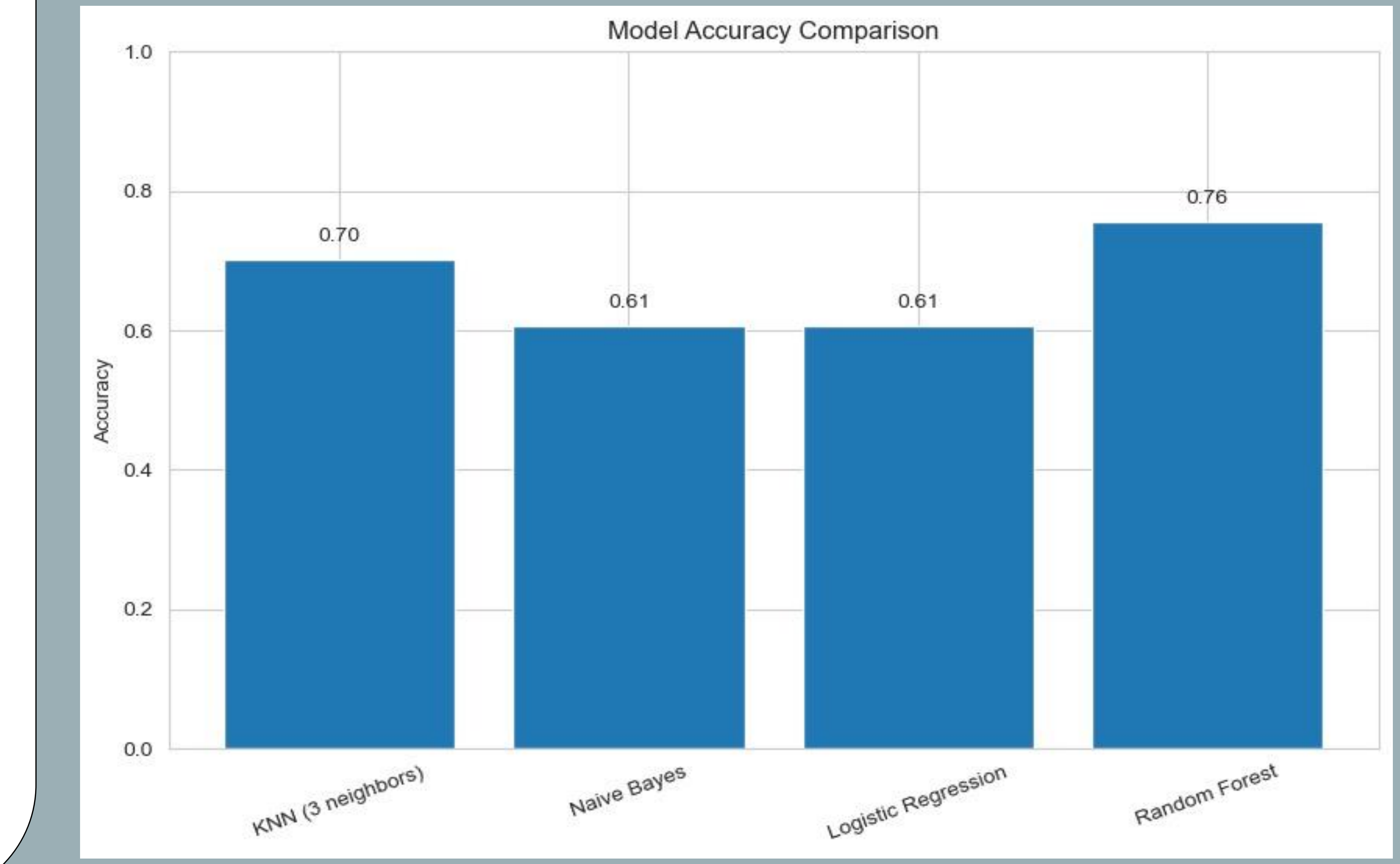
- Data overview: The dataset consisted of approximately 1500 web-scraped social media posts, each labeled with a specific emotional sentiment. Initial exploration revealed over 100 unique emotion labels, many of which were duplicated or inconsistently formatted. To address this, labels were cleaned by removing whitespace, standardizing capitalization, and grouping similar emotions into four broader categories: positive, negative, neutral, and other.
- Text data was vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) with the top 1,000 features selected. This preprocessing step converted text into numerical feature vectors suitable for machine learning models.
- The cleaned dataset was split into 80% training and 20% testing sets to evaluate model performance.
- Four classification models were trained and compared: K-Nearest Neighbors (KNN), Multinomial Naive Bayes, Logistic Regression, and Random Forest.
- Model performance was evaluated using accuracy scores, confusion matrices, and learning curves. Cross-validation was applied during learning curve generation to assess generalization.

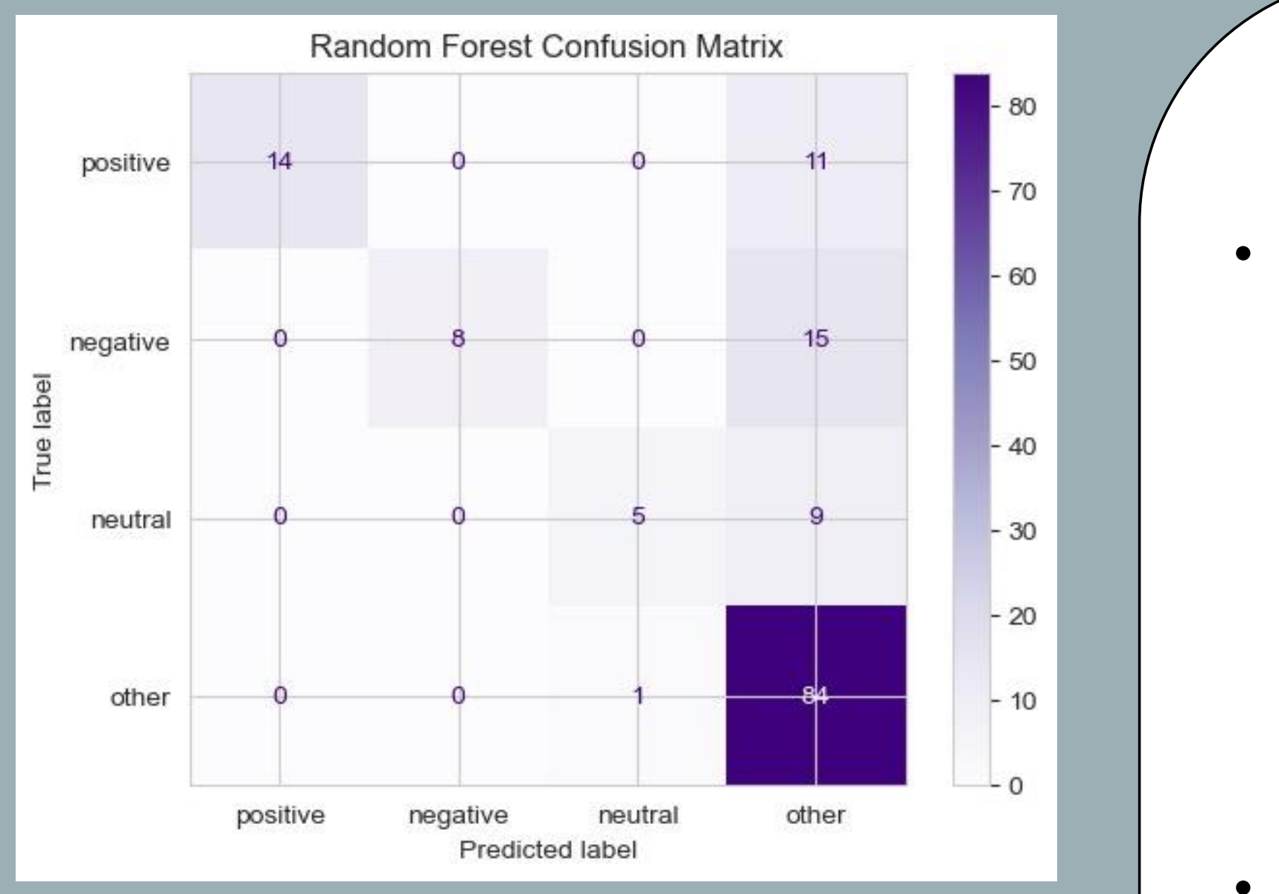


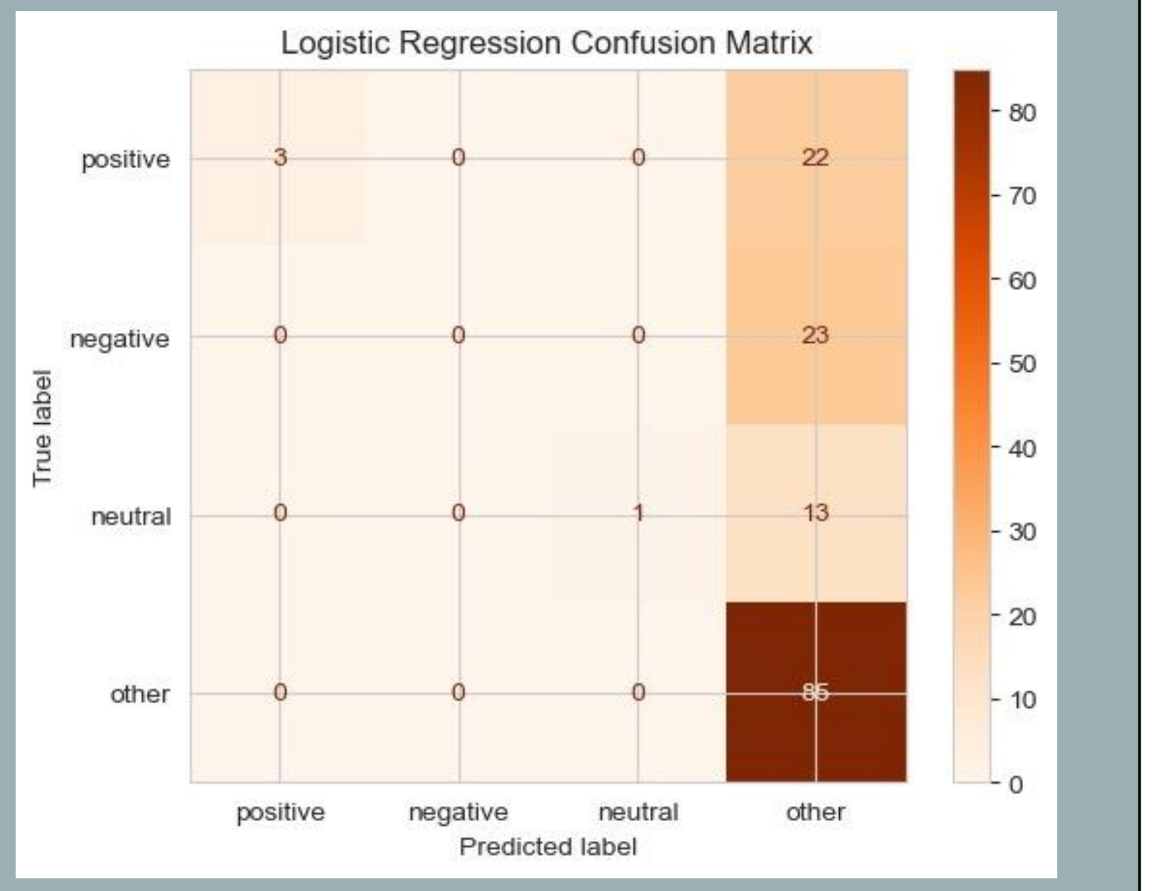


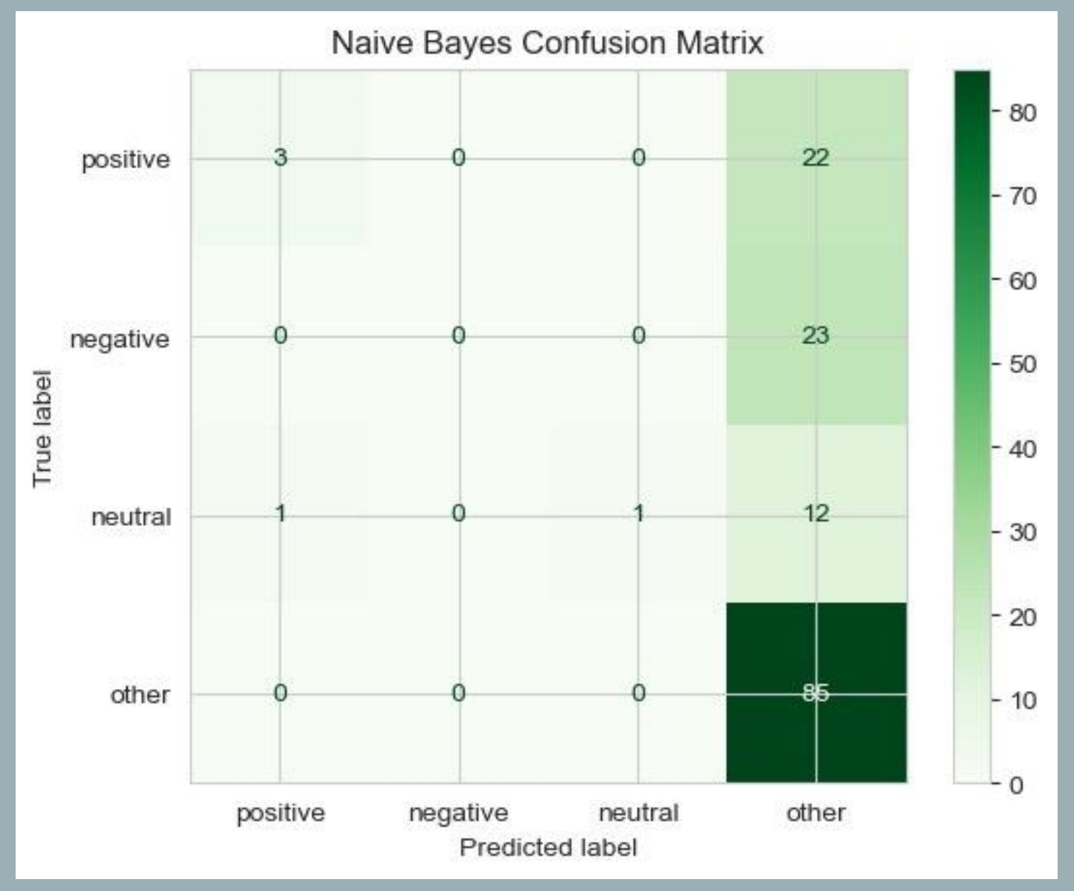


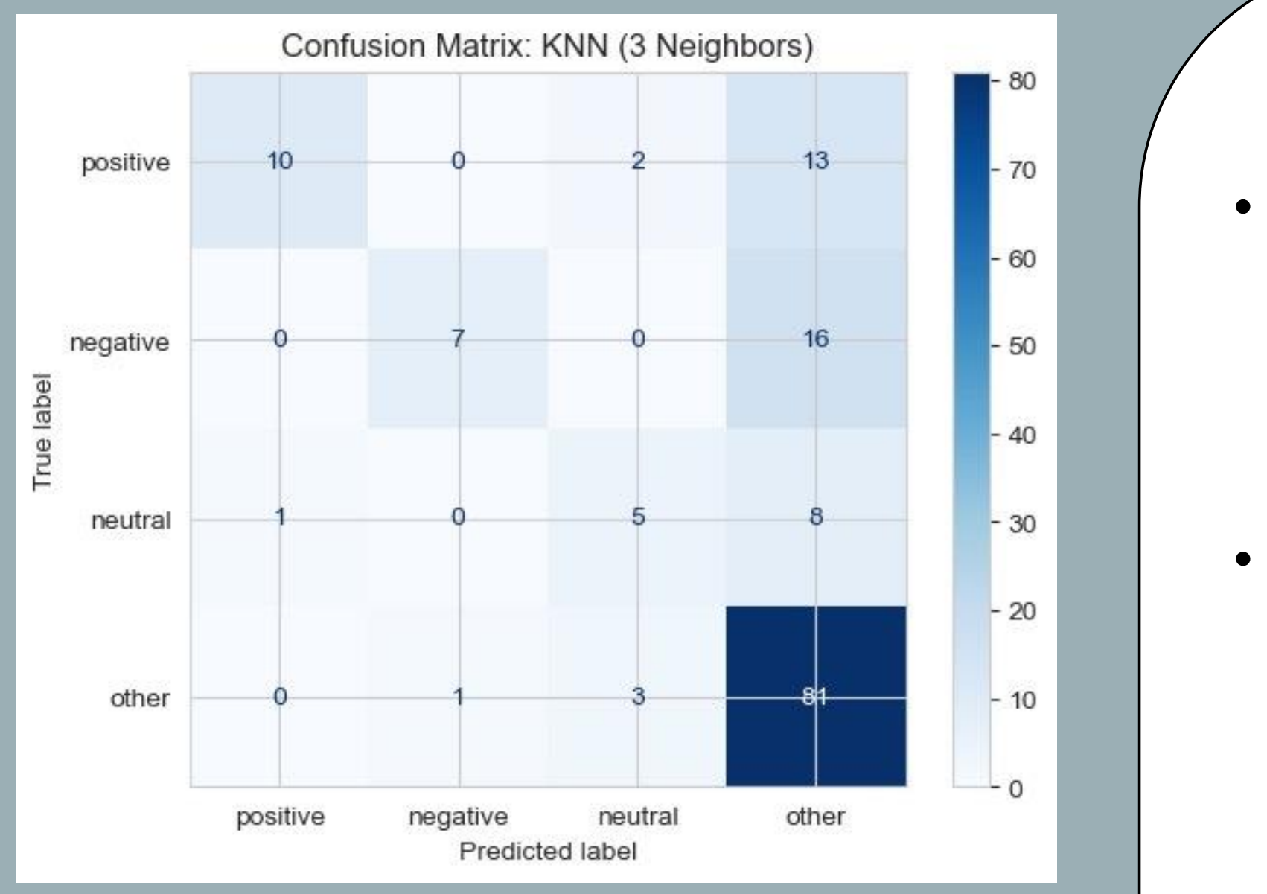


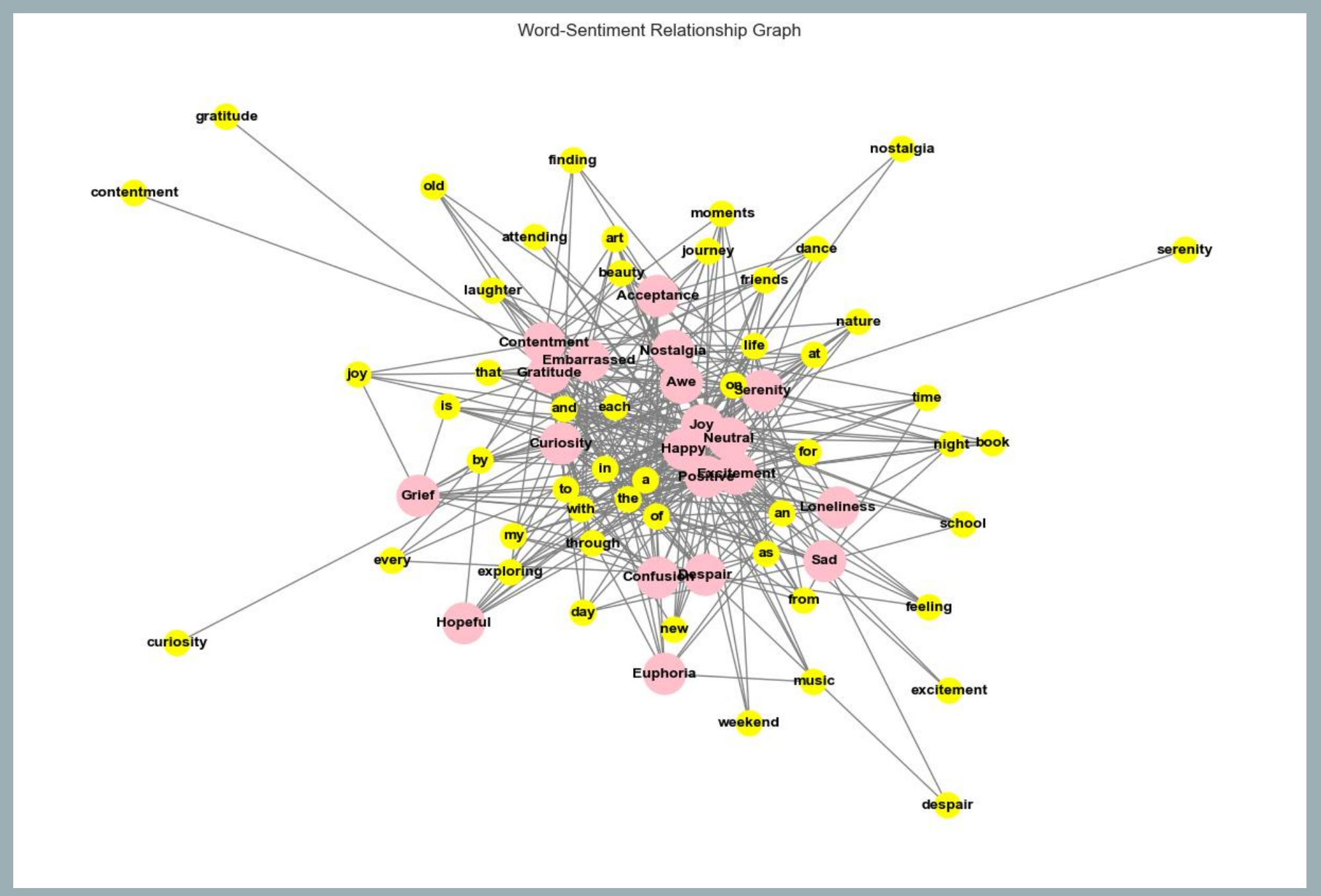












DISCUSSION

- K-Nearest Neighbors (KNN - 3 Neighbors): Reducing the number of neighbors to 3 slightly improved model flexibility but did not dramatically boost accuracy. While KNN (3 neighbors) maintained similar testing performance, it provided more detailed local fits. However, confusion between "other" and emotional categories remained, indicating challenges in fine-grained classification.
- Naive Bayes: Naive Bayes achieved a testing accuracy of 61%, with a strong bias toward predicting the "other" category. While the model performed consistently across training sizes, it struggled with subtle emotional differences in text data. Its simplicity limited its effectiveness in capturing complex sentiment patterns.
- Logistic Regression: Logistic Regression also reached an accuracy of 61%, mirroring Naive Bayes in overall behavior. The model provided a solid linear baseline but failed to separate nuanced emotions effectively, often defaulting to "other" as the prediction. This suggests the need for more complex modeling for text-based emotion detection.
- Random Forest Classifier: Random Forest was the highest-performing model, achieving a testing accuracy of 76%. It balanced precision and recall across the major sentiment groups and demonstrated the best ability to capture complex word usage patterns. Despite a slight bias toward "other" predictions, Random Forest generalized well and showed the most robust performance across learning curves.

CONCLUSION

- Word usage patterns are strongly correlated with emotional tone in social media posts, supporting the value of deeper sentiment analysis beyond basic positive/negative classification.
- Random Forest was the most effective model, achieving the highest testing accuracy (~76%) and best generalization across emotions compared to simpler models like Naive Bayes and Logistic Regression.
- Grouping emotions into broader categories (positive, negative, neutral, other) significantly improved model performance and reduced noise from highly specific labels.
- K-Nearest Neighbors hyperparameter tuning (reducing neighbors from 5 to 3) demonstrated that model flexibility can impact performance, although Random Forest still outperformed KNN overall.
- Future work should focus on expanding the dataset size, refining emotion categories, and exploring deep learning approaches (e.g., LSTM, BERT) for even better sentiment detection.