Sentiment Analysis on Social Media Data

Analyzing Public Opinion Using Machine Learning Models

Name – Sumit Baviskar
Date – 14 Jan 2025

Overview

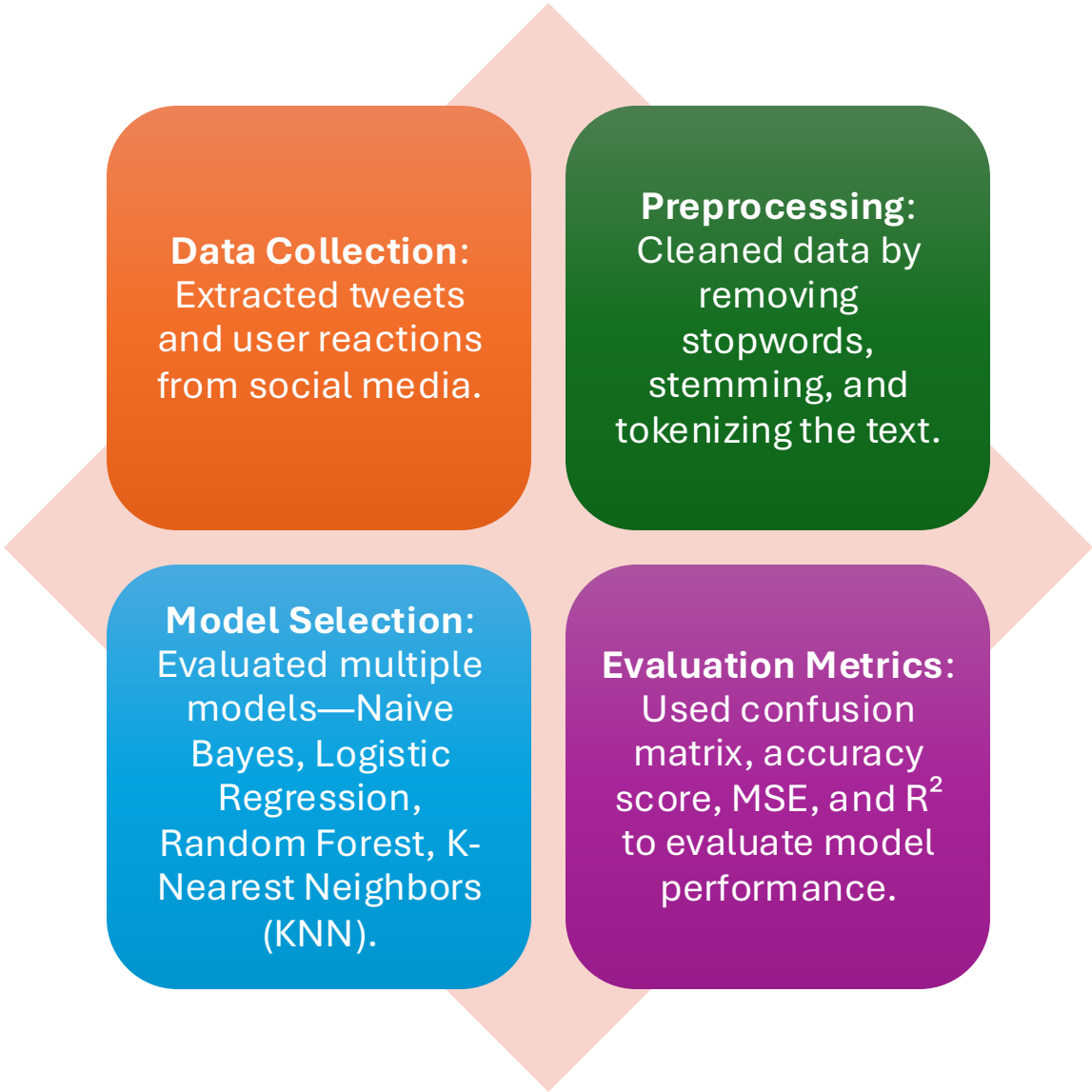
Analyze sentiment (positive, negative, neutral, irrelevant) in social media posts (e.g., Twitter).

Understand public opinion trends, brand perception, and campaign impact.

Problem Statement

- **Problem**: Social media platforms generate vast amounts of data that express public sentiment. Analyzing this unstructured data to extract actionable insights is challenging.
- **Goal**: Build a sentiment analysis model that can efficiently classify social media posts into different sentiment categories (positive, negative, neutral, irrelevant).

Methodology



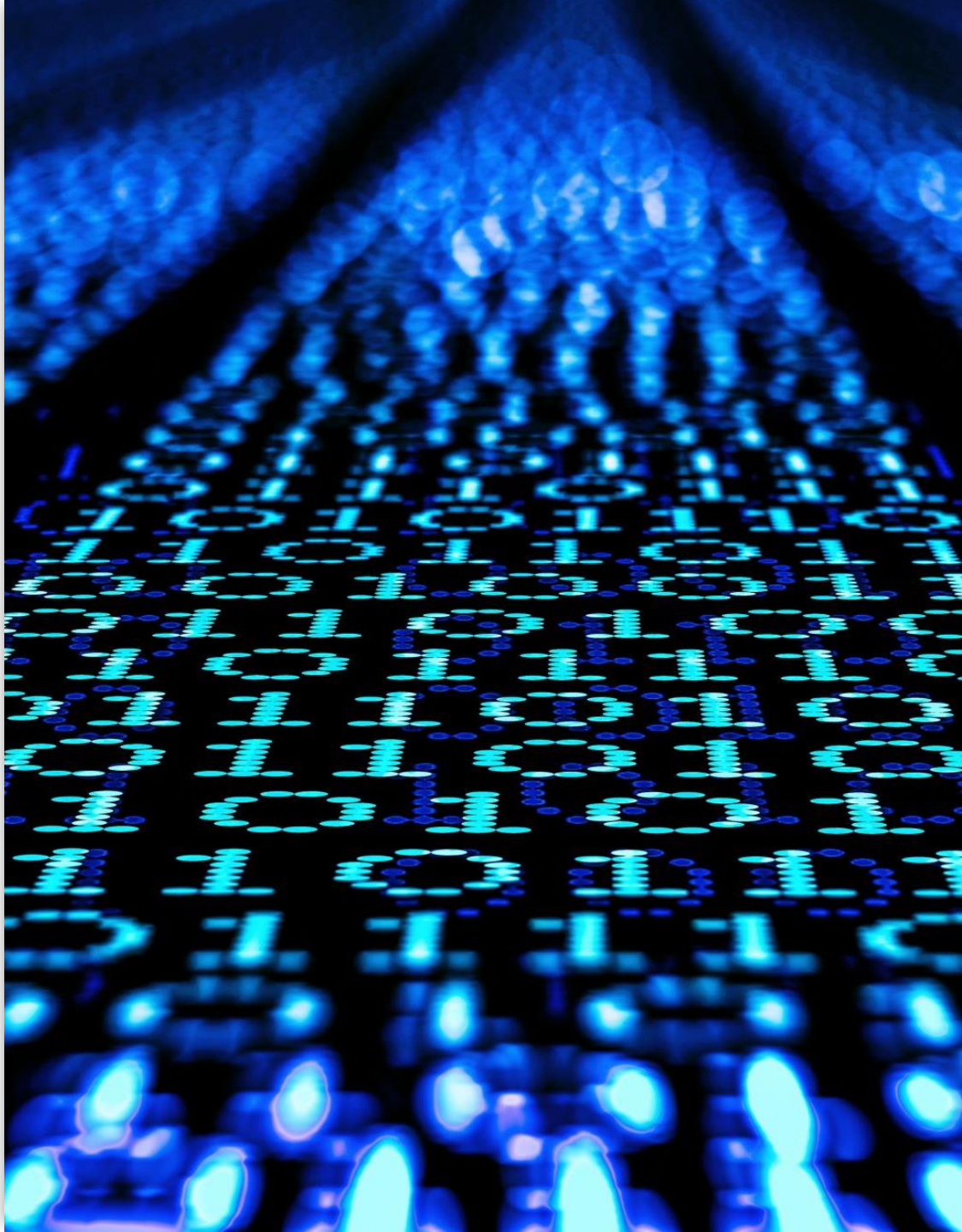
Data Preprocessing

Text Cleaning:

- Removed non-alphabetical characters.
- Lowercased text and split into words.
- Removed stopwords and applied stemming.

Feature Extraction:

- Converted text into numerical data using CountVectorizer.



Model Evaluation - Results

- **Random Forest** achieved the highest accuracy and lowest MSE, making it the most effective model.
- **KNN** performed well with accuracy above 80%.
- **Logistic Regression** and **Naive Bayes** underperformed in comparison.

	Model	R2_Score	Accuracy_Score	MSE
0	Naive Bayes	-0.122708	0.437164	1.584118
1	Logistic Regression	-0.072859	0.583839	1.513781
2	Random Forest	0.584832	0.844184	0.585793
3	K-Neighbors	0.468566	0.807341	0.749843

Recommendation

01

Best Model: Use **Random Forest** for sentiment analysis due to its superior performanc

02

Experiment with hyperparameter tuning for **Random Forest** and **KNN**.

03

Explore advanced features like TF-IDF, word embeddings, or ensemble models (e.g., XGBoost).

04

Additional Data: Consider adding more labeled data for training to improve model performance.

Conclusion

Summary: The project demonstrated how sentiment analysis can be used to analyze social media data. The Random Forest model was the most effective in classifying sentiments, achieving an accuracy of 84.22%.

Future Work: Focus on enhancing feature extraction methods and exploring more complex models like deep learning for text classification.

Thank You

Questions? Open for Discussion