Shivam Raj

CSE,2301201

ML-mini project: Design a sentiment analyzer using ML

Sentiment Analysis Project Report

1. Executive Summary

This project implements a Sentiment Analysis pipeline on the 'twitter_training.csv' dataset. The objective was to classify tweets into sentiments (Positive, Negative, Neutral, Irrelevant) using Logistic Regression, Naive Bayes, and K-Nearest Neighbors (KNN).

Result: The KNN model demonstrated superior performance with an Accuracy of 85.2%.

2. Methodology & Logic

A. Data Preprocessing

1. Cleaning: Dropping missing values.
2. Vectorization (TF-IDF): Converting text into numerical vectors, highlighting unique words while downplaying common ones.

B. Model Selection

- Logistic Regression: Linear model using a sigmoid function.
- Naive Bayes: Probabilistic classifier assuming word independence.
- KNN: Classification based on the majority sentiment of 'k' nearest neighbors.

3. Mathematical Framework

The following formulas were utilized to calculate performance:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Algorithm Logic:

TF-IDF: weight = tf * log(N / df)

Logistic Sigmoid: P(y=1|x) = 1 / (1 + e^-(w^T x + b))

KNN Euclidean Distance: d(x, y) = sqrt(sum((xi - yi)^2))

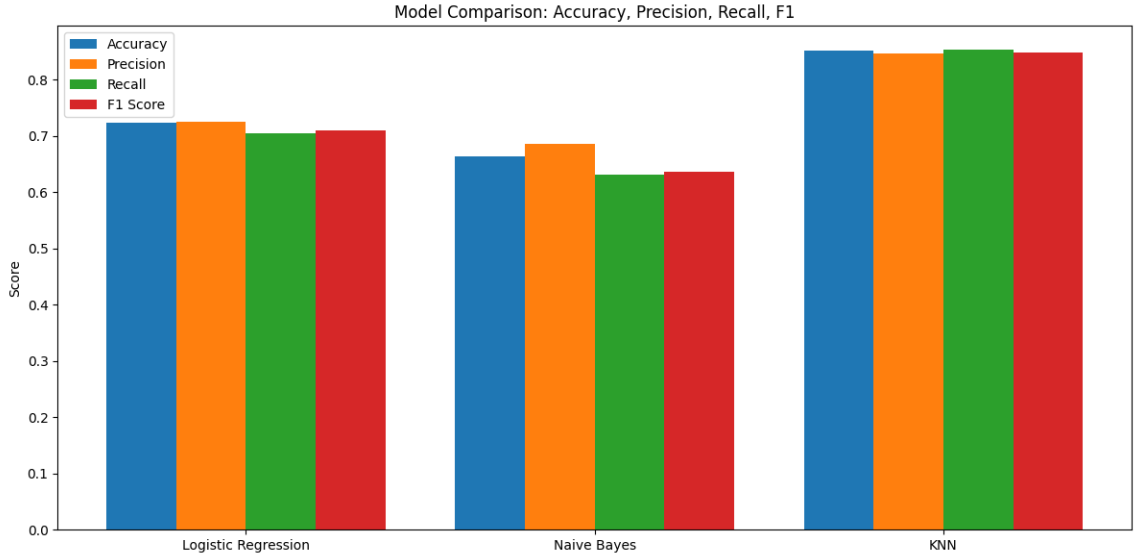
4. Quantitative Results

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7230	0.7247	0.7042	0.7103
Naive Bayes	0.6633	0.6866	0.6304	0.6367
KNN	**0.8521**	**0.8461**	**0.8528**	**0.8473**

Observation: KNN is the clear winner, performing significantly better than the baseline models.

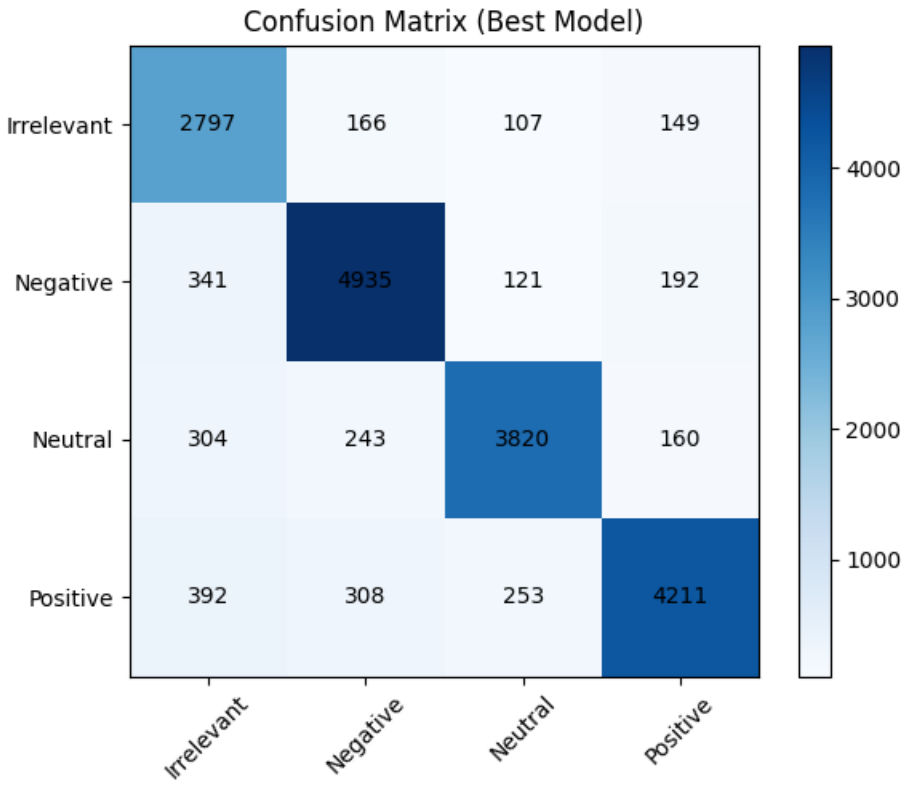
5. Visual Analysis

A. Model Comparison



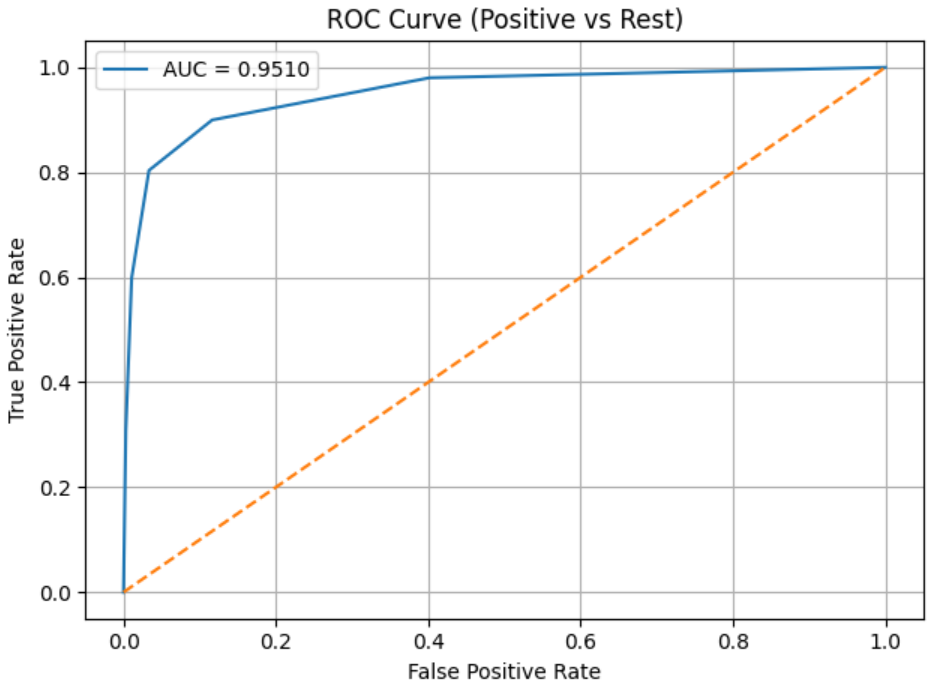
Analysis: The bar chart shows KNN (Right) dominating across all metrics compared to Naive Bayes and Logistic Regression.

B. Confusion Matrix (Best Model)



Analysis: The dark diagonal indicates high correct predictions. The model is strongest at detecting Negative sentiments (4935 correct).

C. ROC Curve



Analysis: AUC = 0.9510. The curve hugs the top-left corner, indicating excellent capability in distinguishing Positive tweets from others.

6. Conclusion

The experiment concludes that KNN is the most effective model for this dataset (F1 ~0.85). The high AUC of 0.95 confirms the model's reliability.