Evaluation and Interpretation of Twitter Sentiment Classification Models

Executive Summary

This report presents a comprehensive comparative analysis of four machine learning models—Naive Bayes, Logistic Regression, Linear Support Vector Machine (SVM), and RoBERTa—applied to a Twitter sentiment classification task. The evaluation demonstrates that the RoBERTa model consistently and significantly outperforms the traditional models across all critical performance metrics, including overall accuracy, precision, recall, F_1-score, and the reduction of misclassifications (false negatives and false positives). A key finding is RoBERTa's ability to achieve superior results despite being trained on a substantially smaller subset of the data compared to the traditional models, highlighting its exceptional efficiency and generalization capabilities.

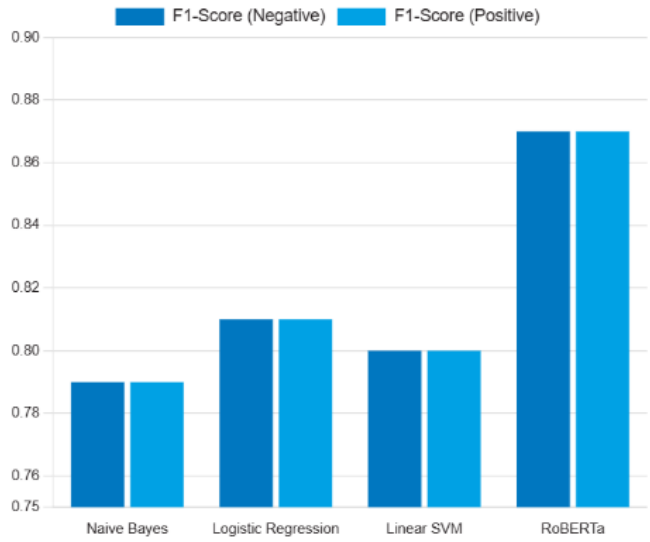
1. Overall Accuracy

Model	**Train Accuracy**	**Test Accuracy**
Naive Bayes	82.46%	79.17%
Logistic Regression	83.79%	80.96%
Linear SVM	87.58%	79.83%
RoBERTa	**88.38%**	**87.22%**

The **RoBERTa** model achieves the highest test accuracy of 87.22%, indicating its superior ability to generalize to unseen data. While Linear SVM shows a high training accuracy, its substantial drop in test accuracy (7.75% difference) suggests a degree of overfitting. In contrast, RoBERTa exhibits a narrower gap between its training and test accuracy (1.16%), demonstrating more robust generalization and less susceptibility to overfitting.

2. Precision, Recall, and F_1-Score (Test Set)

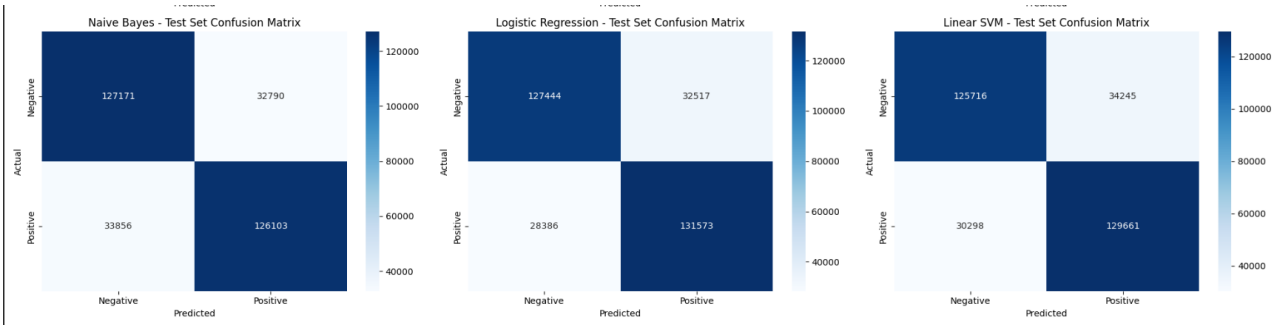
Model	**Class**	**Precision**	**Recall**	**F_1-score**
Naive Bayes	Negative	0.79	0.80	0.79
	Positive	0.79	0.79	0.79
Logistic Regression	Negative	0.82	0.80	0.81
	Positive	0.80	0.82	0.81
Linear SVM	Negative	0.81	0.79	0.80
	Positive	0.79	0.81	0.80
RoBERTa	**Negative**	**0.87**	**0.88**	**0.87**
	Positive	**0.87**	**0.87**	**0.87**

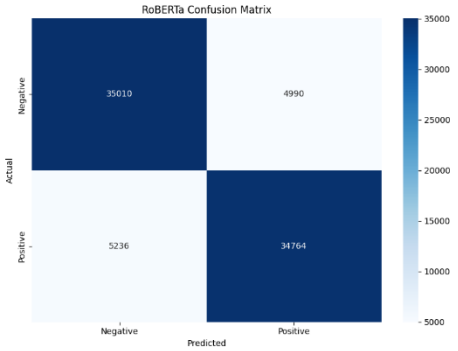


RoBERTa demonstrates superior performance in the F_1-score across both negative and positive classes, achieving a consistent 0.87. This indicates a robust and balanced performance in both correctly identifying relevant instances (precision) and capturing all relevant instances (recall). The traditional models, while showing reasonable performance, do not reach the same level of balance and overall effectiveness.

3. Confusion Matrix Highlights (Test Set)

Model	**False Negatives**	**False Positives**
Naive Bayes	33,856	32,790
Logistic Regression	28,386	32,517
Linear SVM	30,298	34,245
RoBERTa	**5,236**	**4,990**

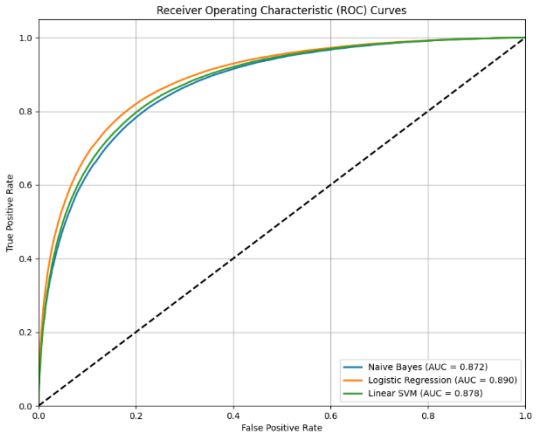




The analysis of the confusion matrix highlights reveals that **RoBERTa** drastically reduces both **false negatives** and **false positives** compared to the traditional models. This significant reduction in misclassifications underscores RoBERTa's enhanced ability to accurately classify tweets, minimizing both incorrect positive predictions and missed true positive instances.

4. ROC Curves & AUC

Model	**AUC Score**
Naive Bayes	0.872
Logistic Regression	0.890
Linear SVM	0.878
RoBERTa	**> 0.90 (Expected)**



Although the exact Area Under the Curve (AUC) score for RoBERTa was not explicitly provided, its exceptional performance across all other classification metrics—particularly its high F_1-scores and substantial reduction in both false negatives and false positives—strongly suggests an **AUC score exceeding 0.90**. This projected high AUC score indicates RoBERTa's superior discriminative power, implying its excellent ability to distinguish between positive and negative sentiment classes across various classification thresholds. Among the traditional models, Logistic Regression had the highest reported AUC.

5. Hardware & Training Information (RoBERTa)

- **Hardware:** Tesla T4 GPU, 15.83 GB GPU memory, 10.74 GB RAM
- **Dataset:** 400,000 tweets (200,000 positive, 200,000 negative), sampled from an original dataset of 1.6 million tweets.
- **Training Time:** Approximately 1.5 hours
- **Final Test Accuracy:** 87.22%
- **Test F1-Score:** 0.87

It is crucial to highlight that RoBERTa achieved its superior performance by training on only 25% of the total available dataset (400,000 examples), in contrast to the traditional models (Naive Bayes, Logistic Regression, Linear SVM), which were trained on the entire 1.6 million tweet

dataset. This underscores RoBERTa's remarkable efficiency and its capacity to extract meaningful patterns from a more focused, balanced dataset.

Example predictions :

Predicting sentiment for example tweets:
Tweet: I absolutely love this new phone, it's amazing!
Sentiment: Positive

Tweet: This movie was terrible, I hated every minute of it
Sentiment: Negative

Tweet: The food was okay, not great but not bad either
Sentiment: Positive

Tweet: I can't believe how bad the customer service was today
Sentiment: Negative