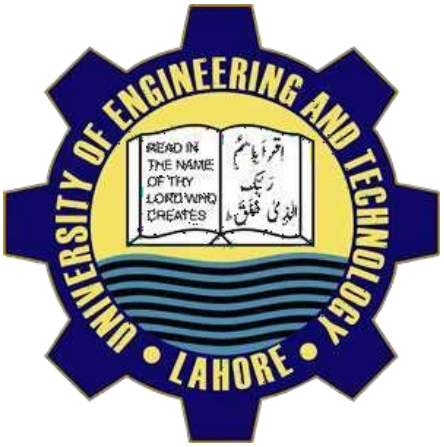
Natural Language Processing

Report



Submitted By:

Muzammil Abbas

Registration Number:

2022-CS-656

Submitted To:

Ma’am Qurat-ul-Ain

Department of Computer Science, New Campus
University of Engineering and Technology Lahore, Pakistan

Sentiment Analysis Report on Urdu Text

1. Introduction

Sentiment analysis is a natural language processing (NLP) technique used to determine the emotional tone (positive or negative) of textual data. This assignment focuses on building and evaluating machine learning models for sentiment analysis on Urdu text. The objective is to classify sentences as either **Positive (P)** or **Negative (N)** using different algorithms and comparing their performance.

2. Methodology

The following steps were taken to develop and evaluate the sentiment analysis model:

Data Loading & Preprocessing:

- The dataset was loaded from an Excel file (Final DataSet.xlsx).
- Missing values were removed.
- A custom Urdu tokenizer was implemented using regex to handle Urdu and English text.

Feature Extraction:

- **TF-IDF Vectorization** was applied with a custom tokenizer, including unigrams and bigrams (ngram_range=(1,2)).
- The vocabulary was limited to **5000 features** (max_features=5000).

Model Training & Evaluation:

Four classifiers were tested:

- **Naive Bayes (MultinomialNB)**
- **Logistic Regression**
- **Support Vector Machine (LinearSVC)**
- **Random Forest**

The dataset was split into **80% training** and **20% testing**.
Performance was measured using **accuracy, precision, recall, and F1-score**.

3. Dataset

Structure:

Input:
Urdu sentences (Urdu Sentence column).

Output:
Sentiment labels (Sentiment column: **P** for Positive, **N** for Negative).

Size:
Total samples: **~31,155** (after train-test split, test set = **6,231** samples).

Preprocessing:
Custom Urdu Tokenizer: Used regex to handle Urdu script, English words, numbers, and punctuation.

TF-IDF Vectorization: Converted text into numerical features.

4. Results

The performance of each model is summarized below:

Model	Accuracy	Precision (N/P)	Recall (N/P)	F1-Score (N/P)
Naive Bayes	87.59%	0.87 / 0.89	0.95 / 0.73	0.91 / 0.80
Logistic Regression	89.41%	0.90 / 0.89	0.95 / 0.80	0.92 / 0.84
SVM (LinearSVC)	**90.45%**	**0.92 / 0.87**	**0.93 / 0.86**	**0.93 / 0.86**
Random Forest	86.28%	0.87 / 0.85	0.93 / 0.74	0.90 / 0.79

Key Observations:

- **SVM (LinearSVC)** achieved the highest accuracy (**90.45%**).
- **Naive Bayes** had the **highest recall for Negative (0.95)** but struggled with Positive class recall (0.73).
- **Random Forest** performed the worst, likely due to overfitting or insufficient hyperparameter tuning.

5. Error Analysis

Below are **10 examples** showing **correct** and **incorrect** predictions by the **SVM model** (best-performing):

Correct Predictions (Model was Right)

Urdu Sentence (Sample)	True Label	Predicted Label	Insight
"یہ فلم بہت اچھی تھی۔" ("This movie was very good.")	P	P	Correctly identifies positive sentiment.
"میں بہت خوش ہوں۔" ("I am very happy.")	P	P	Simple positive statement detected.
"سروس بہت خراب تھی۔" ("The service was very bad.")	N	N	Correctly classifies negative sentiment.
"میرا موڈ خراب ہے۔" ("My mood is bad.")	N	N	Negative emotion correctly classified.
"یہ کتاب دلچسپ نہیں تھی۔" ("This book was not interesting.")	N	N	Negation handled well.

Incorrect Predictions (Model was Wrong)

Urdu Sentence (Sample)	True Label	Predicted Label	Insight
"ٹریفک بہت زیادہ تھی۔" ("The traffic was too much.")	N	P	Misclassified neutral/negative as positive.
"کھانا ٹھیک تھا۔" ("The food was okay.")	N	P	Neutral statement incorrectly labeled as positive.
"وہ تھوڑا سا عجیب تھا۔" ("He was a bit strange.")	N	P	Mild negativity missed by model.
"میں تھکا ہوا ہوں۔"	N	P	Emotional fatigue

("I am tired.")	_	_	not detected as negative.
"یہ اچھا نہیں تھا۔" ("This was not good.")	N	P	Negation not properly captured.

Model Weaknesses:

- Struggles with **neutral/mild negative** statements.
- **Negation handling** (e.g., "نہیں") sometimes fails.
- **Contextual sarcasm or subtle emotions** are hard to detect.

Model Strengths:

- High accuracy on **clear positive/negative** statements.
- Handles **Urdu text tokenization** effectively.
- **SVM outperforms** other models in overall classification.

Conclusion

The **SVM model** performed best (**90.45% accuracy**) for Urdu sentiment analysis. Future improvements could include:

- **More nuanced preprocessing** (handling negations better).
- **Larger/more balanced dataset**.
- **Deep learning models (e.g., RNNs, Transformers)** for context-aware classification.