Report on Urdu Sentiment Analysis

# 1. Evaluation

## Model Performance

The sentiment analysis model achieved an accuracy of approximately 78.61% on the validation set of Urdu posts. The detailed classification report is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0.0 | 0.78 | 0.78 | 0.78 | 1831 |
| 1.0 | 0.80 | 0.79 | 0.79 | 1994 |
| Accuracy |  |  | 0.79 | 3825 |
| Macro Avg | 0.79 | 0.79 | 0.79 | 3825 |
| Weighted Avg | 0.79 | 0.79 | 0.79 | 3825 |

## Analysis of Performance

### Strong Points:

• The model performs relatively well on both classes, achieving a precision of 80% for class 1.0 (sarcastic) and 78% for class 0.0 (not sarcastic).

• The F1-scores are balanced, indicating a good trade-off between precision and recall.

### Weak Points:

• The model struggles with complex sentences, particularly those containing nuanced sarcasm or cultural references that are common in colloquial Urdu.

• Instances of overlapping sentiment expressions may lead to misclassifications, especially in posts with mixed sentiments.

### Areas for Improvement

• Data Quality: Enhance the dataset by collecting more examples of complex and sarcastic sentences. Cleaning and preprocessing steps could be optimized further to handle colloquial expressions better.

• Feature Engineering: Explore additional features such as syntactic structures, sentiment lexicons specific to Urdu, or even multilingual embeddings to improve understanding.

• Model Selection: Experiment with different models or ensemble methods that can better capture the nuances in language.

# 2. Challenges in Urdu Sentiment Analysis

## Complex Morphology

Urdu has a rich morphological structure with inflections and derivations that can lead to multiple forms of a single word. This complexity can make it difficult for models to capture the intended sentiment accurately.

## Colloquial Language

Social media posts often contain slang, abbreviations, and informal expressions, which may not be represented well in traditional language models. This can result in poor generalization on unseen data.

## Noisy Data

Data collected from social media is often noisy, containing spelling mistakes, informal language, and extraneous elements (like emojis or hashtags). Cleaning this data effectively while retaining useful sentiment information is a significant challenge.

## Contextual Understanding

Sarcasm and irony heavily rely on context. The model's ability to understand the subtleties and implicit meanings in sentences can greatly influence its performance, particularly in identifying sarcastic remarks.