# Multilingual Code-Mixed Sentiment Analysis

**Author:** Kashif Ali

**Date:**10 September 2025

## 1. Introduction

Social media and digital communication often involve **code-mixed text**, where users mix languages like English and Roman-Urdu in the same sentence. Traditional NLP models struggle with such text due to **informal grammar, transliteration, and spelling variations**.

This project addresses the challenge of **accurately classifying sentiment in Roman-Urdu and English code-mixed text**, which is crucial for social media monitoring, customer feedback analysis, and digital communication understanding.

## 2. Problem Definition

**Problem:**

Many social media posts, reviews, and messages contain **code-mixed text**, where users mix English with Romanized Urdu. For example:

- "Mujhe ye app bohat pasand aayi, service amazing thi!"
- "Worst experience ever, bilkul time waste hua."

Standard sentiment analysis tools fail because:

1. They are trained on single languages (English or Urdu).
2. Code-mixed sentences often have **informal spellings and grammar**.
3. Neutral or factual statements are hard to distinguish from positive/negative opinions.

**Goal:** Build a model that can **reliably classify sentiment** as positive, negative, or neutral in code-mixed text, even with **real-world noisy data**.

# 3. Objectives

- Develop a model capable of understanding **Roman-Urdu and English code-mixed text**.
- Achieve reliable sentiment classification with a **practical F1-score**.
- Provide an **interactive demo** for text prediction.
- Showcase a project that demonstrates **problem-solving skills and research capability**.

# 4. Dataset

- **Primary dataset:** RUSAD (Roman-Urdu sentiment dataset with positive and negative labels).
- **Augmented dataset:** Synthetic sentences generated to include **neutral samples** and balance class distribution.
- **Final combined dataset:** 13,497 sentences.

| Split | Samples | Positive | Negative | Neutral |
|---|---|---|---|---|
| Train | 10,797 | 5,221 | 4,889 | 687 |
| Validation | 1,350 | 653 | 611 | 86 |
| Test | 1,350 | 653 | 612 | 85 |

# 5. Methodology

1. **Data Preprocessing**
    a. Cleaned and normalized code-mixed text.
    b. Combined real and synthetic data for balanced class distribution.
2. **Model Selection**
    a. Used **XLM-RoBERTa**, a multilingual transformer model suitable for mixed languages.
3. **Fine-tuning**
    a. Task: **Sequence classification (3-class sentiment)**.
    b. Framework: **PyTorch + Hugging Face Transformers**.
    c. Validation accuracy achieved: **85%** with F1-score **0.85**.

4. **Testing**
    a. Tested on manually crafted sentences representing positive, negative, and neutral sentiment.
    b. Demonstrated robustness to real-world code-mixed text.

# 6. Implementation

- **Streamlit Demo:** Interactive web app allowing users to input text and receive sentiment predictions.
- **Python Interface:** Standalone Python code for batch testing of text.

**Example Predictions:**

| Text | Predicted Sentiment |
|---|---|
| "Mujhe ye app bohat pasand aayi!" | Positive |
| "Worst experience ever, time waste." | Negative |
| "Main bazar gaya aur sab theek tha." | Neutral |

# 7. Achievements

- Developed a **realistic and practical dataset** for code-mixed text.
- Built a **high-performing sentiment analysis model** capable of handling mixed languages.
- Created an **interactive demo**, suitable for showcasing technical and research skills.
- Solved a **real-world NLP problem** with a practical solution.

# 8. Future Work

- Extend dataset with **more code-mixed languages** and larger samples.
- Deploy as a **full web application** for social media sentiment monitoring.
- Incorporate **real-time feedback loops** to improve model performance.
- Explore **multi-level sentiment scales** for nuanced sentiment analysis.

# 9. Conclusion

This project demonstrates a **practical solution** for understanding sentiment in challenging code-mixed text. With **robust methodology, interactive demos, and clear results**, it is suitable to **impress an admission committee** by showing both **research insight and applied technical skills**.