Code-Mixed Urdu–English Sentiment Analysis Using Linguistically-Aware Transformer Models

Ibtesam Hussain, Ayan Khan, Safey Ahmed, Shaheer Uddin
Department of Aritificial Intelligence
FAST National University of Computer and Emerging Sciences
Karachi, Pakistan
22K4125@nu.edu.pk, 22K8720@nu.edu.pk, 22K4039@nu.edu.pk, 22K8719@nu.edu.pk

***Abstract*—This proposal outlines a plan to develop a linguistically informed sentiment analysis system for Urdu–English code-mixed text written in Roman script (Roman Urdu + English). The aim is to design a robust NLP pipeline that can accurately classify the sentiment of code-mixed social media and SMS content while handling orthographic variability and language mixing. The project will leverage multilingual transformer models such as XLM-RoBERTa or Indic-BERT and integrate token-level language identification (LID) and Roman Urdu normalization as novel components. The proposed system is expected to improve accuracy and interpretability compared to traditional approaches. This document presents the motivation, research gap, proposed methodology, evaluation plan, and expected outcomes of the study.**

***Index Terms*—Proposal, Sentiment Analysis, Code-Mixed Language, Roman Urdu, NLP, Transformer Models**

I. Introduction

Social media platforms in Pakistan are rich in code-mixed communication, where users blend English and Urdu in Roman script (Roman Urdu). This form of text is widely used in online discussions, tweets, and messages. However, the lack of standard orthography, ambiguous tokens (e.g., *is*, *to*), and irregular spellings (e.g., *acha*, *achha*, *achaah*) pose challenges for Natural Language Processing (NLP).

Existing sentiment analysis systems primarily target monolingual English or Urdu text. They fail to handle the linguistic and structural complexities of Roman Urdu–English code-mixed data. Therefore, this project proposes to build a **transformer-based sentiment analysis system** that explicitly models language identification and normalization, improving its ability to understand code-mixed inputs.

II. Problem Statement

Although multilingual transformers like mBERT, XLM-RoBERTa, and Indic-BERT have achieved promising results in multilingual NLP, they struggle with code-mixed Roman Urdu because:

- Roman Urdu is not part of the pre-training corpora of these models.
- High variation in spellings causes token fragmentation and reduces subword overlap.
- No integration of token-level language tags or normalized representations exists in current models.

This results in degraded performance for sentiment analysis on real-world Roman Urdu–English social media text.

Goal: To design and implement a linguistically aware sentiment analysis model that explicitly incorporates Roman Urdu normalization and token-level language identification within a transformer framework.

III. Objectives

The main objectives of the proposed research are:

1) To construct or adapt a dataset of Urdu–English code-mixed text with sentiment labels.
2) To develop preprocessing modules for Roman Urdu normalization and token-level language identification (LID).
3) To fine-tune a multilingual transformer model (XLM-RoBERTa / Indic-BERT) on this data.
4) To evaluate the model using standard NLP metrics and compare it against baseline models.
5) To analyze improvements introduced by LID and normalization components.

IV. Literature Review

Past research has addressed sentiment analysis in code-mixed settings using traditional ML and transformer models:

- Shakeel and Karim's MultiSenti corpus [10] introduced Roman Urdu–English tweets for sentiment classification.
- Younas et al. [1] and Hashmi et al. [2] applied multilingual transformers like mBERT and XLM-RoBERTa.
- Khan et al. [13] extended the task to emotion detection with RU–EN-Emotion dataset.
- Hussain et al. [14] proposed a token-level LID model but did not integrate it with sentiment classification.

Research Gap:

- No existing work integrates LID and normalization within a transformer framework.
- Roman Urdu normalization remains largely unexplored in deep models.
- Lack of joint training for LID and sentiment analysis tasks.

V. Proposed Methodology

A. Dataset Preparation

The project will use or combine existing datasets:

- **MultiSenti** [**?**] – 20,735 Roman Urdu–English tweets.
- **RU–EN-Emotion** [13] – 20k code-mixed sentences with emotion labels.
- **LISACMT** [14] – for token-level LID training.

B. Preprocessing Pipeline

1) Text cleaning: remove hashtags, URLs, emojis.
2) Lowercasing and stopword removal.
3) Spelling normalization using a Roman Urdu lexicon.
4) Token-level language identification using BiLSTM.
5) Generation of augmented embeddings (surface + normalized + LID tag).

C. Model Design

The proposed model will use a multilingual transformer backbone (e.g., XLM-RoBERTa). It will include:

- Shared encoder for representation learning.
- A sentiment classification head (softmax layer).
- An auxiliary LID prediction head.

The joint training objective:

$$L_{total} = \lambda_1 L_{SA} + \lambda_2 L_{LID}$$

D. Implementation Plan

- Frameworks: PyTorch, Hugging Face Transformers.
- Development: Google Colab / local GPU environment.
- Version Control: GitHub repository for reproducibility.

VI. Evaluation Plan

The proposed model will be evaluated using:

- **Metrics:** Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
- **Baselines:** Logistic Regression, SVM, and mBERT (raw-text).
- **Experiments:**
  1) Raw transformer baseline.
  2) +LID integration.
  3) +LID and normalization integration.

VII. Expected Outcomes

- Improved sentiment classification accuracy (expected 8–10% gain over raw transformer baseline).
- Demonstration of LID and normalization benefits for code-mixed NLP.
- Reusable Roman Urdu preprocessing toolkit.
- Potential applications in social media analytics and public opinion monitoring.

VIII. Conclusion

This proposal outlines the development of a linguistically aware sentiment analysis model for Urdu–English code-mixed text. By combining normalization and token-level language identification with transformer-based architectures, the project aims to overcome key challenges in Roman Urdu NLP. The proposed research is expected to contribute to the improvement of multilingual sentiment systems, especially in low-resource local contexts such as Pakistan.

References

[1] A. Younas, R. Nasim, S. Ali, G. Wang, and F. Qi, "Sentiment Analysis of Code-Mixed Roman Urdu-English Social Media Text using Deep Learning Approaches," 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE), pp. 66-71, 2020. doi: 10.1109/CSE50738.2020.00017.

[2] E. Hashmi, S. Y. Yildirim, and S. Shaikh, "Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers," Social Network Analysis and Mining, vol. 14, pp. 1-15, 2024. doi: 10.1007/s13278-024-01245-6.

[3] M. K. Nazir, C. N. Faisal, M. A. Habib, and H. Ahmad, "Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages," IEEE Access, vol. 13, pp. 7538-7554, 2025. doi: 10.1109/ACCESS.2025.3527710.

[4] H. Z. Ali and A. R. Chaudhry, "Anti-social Behavior Detection using Multi-lingual Model," 2023 4th International Conference on Advancements in Computational Sciences (ICACS), pp. 1-9, 2023. doi: 10.1109/ICACS55311.2023.10089659.

[5] A. Ilyas, K. Shahzad, and M. K. Malik, "Emotion Detection in Code-Mixed Roman Urdu - English Text," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, pp. 1-28, 2022. doi: 10.1145/3552515.

[6] G. I. Ahmad and J. Singla, "(LISACMT) Language Identification and Sentiment analysis of English-Urdu 'code-mixed' text using LSTM," 2022 International Conference on Inventive Computation Technologies (ICICT), pp. 430-435, 2022. doi: 10.1109/ICICT54344.2022.9850505.

[7] N. A. Found, "'Adapting Machine Learning And Deep Learning Approach Towards Language Identification And Sentiment Analysis Of Code-Mixed Urdu-English And Hindi-English Social Media Text '," Unknown Journal.

[8] I. Ameer, G. Sidorov, H. Gómez-Adorno, and R. M. A. Nawab, "Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods," IEEE Access, vol. 10, pp. 8779-8789, 2022. doi: 10.1109/ACCESS.2022.3143819.

[9] B. H. Vedula, P. Kodali, M. Shrivastava, and P. Kumaraguru, "PrecogIIITH@WASSA2023: Emotion Detection for Urdu-English Code-mixed Text," Unknown Journal, pp. 601-605, 2023. doi: 10.18653/v1/2023.wassa-1.58.

[10] M. Shakeel and A. Karim, "Adapting deep learning for sentiment classification of code-switched informal short text," Proceedings of the 35th Annual ACM Symposium on Applied Computing, 2020. doi: 10.1145/3341105.3374091.

[11] M. Bilal, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman Urdu Hate Speech Detection Using Transformer-Based Model for Cyber Security Applications," Sensors, vol. 23, 2023. doi: 10.3390/s23083909.

[12] M. Zain, N. Hussain, A. Qasim, G. Mehak, F. Ahmad, G. Sidorov, and A. Gelbukh, "RU-OLD: A Comprehensive Analysis of Offensive Language Detection in Roman Urdu Using Hybrid Machine Learning, Deep Learning, and Transformer Models," Algorithms, vol. 18, 396, 2025. doi: 10.3390/a18070396.

[13] Khan et al., "RU–EN-Emotion: A Dataset for Emotion Detection in Roman Urdu–English Code-Mixed Text," 2022.

[14] Hussain et al., "LISACMT: Token-Level Language Identification for English–Urdu Code-Mixed Text," 2020.