Enhancing Urdu Sentiment Classification through Instruction-Tuned LLMs and Cross-Lingual Transfer

Hasan Faraz Khan[1] **Noor Fatima**[1] **Irfan Ahmad**[1,2]
g202427420@kfupm.edu.sa g202427440@kfupm.edu.sa irfan.ahmad@kfupm.edu.sa
[1]Information and Computer Science Department, KFUPM, Dhahran, Saudi Arabia
[2]SDAIA-KFUPM Joint Research Center for AI, Dhahran, Saudi Arabia

Abstract

Sentiment analysis in low-resource languages such as Urdu poses unique challenges due to limited annotated data, morphological complexity, and significant class imbalance in most publicly available datasets. This study addresses these issues through two experimental strategies. First, we explore class imbalance mitigation by using instruction-tuned large language models (LLMs) to generate synthetic negative sentiment samples in Urdu. This augmentation strategy results in a more balanced dataset, which significantly improves the recall and F1-score for minority class predictions when fine-tuned using a multilingual BERT model. Second, we investigate the effectiveness of translating Urdu text into English and applying sentiment classification through a pre-trained English language model. Comparative evaluation reveals that the translation-based pipeline, using a RoBERTa model fine-tuned for English sentiment classification, achieves superior performance across major metrics. Our results suggest that LLM-based augmentation and cross-lingual transfer via translation both serve as viable approaches to overcome data scarcity and performance limitations in sentiment analysis for low-resource languages. The findings highlight the potential applicability of these approaches to other under-resourced linguistic domains.

Keywords: Urdu sentiment analysis; large language models; data augmentation; cross-lingual transfer; machine translation.

1 Introduction

Sentiment analysis has become an important instrument for measuring public opinion, customer attitude, and user-generated content on social media. Although significant advances have been made in languages like English, sentiment analysis in low-resource languages like Urdu is a largely untapped and technically challenging area. The inherent complexities of the Urdu language, such as rich morphology, script variation irregularities, and limited annotated resources, are a few reasons why it is difficult to develop robust sentiment classification systems (Khattak et al., 2021). Furthermore, class imbalance is a common issue in sentiment datasets, where positive or neutral sentiments heavily outweigh negative sentiments, leading to biased learning and poor generalization of minority classes (Ashraf et al., 2023, 2024).

The growing social media and review website presence of Urdu speakers online demands better sentiment analysis tools for this linguistic setting. While there has been growing interest in using deep learning and transformer-based architectures for Urdu and Roman Urdu, most existing work still faces issues with limited resources, class imbalance, and weak generalization (Naqvi et al., 2021; Khan et al., 2021; Altaf et al., 2022). Traditional methods such as Support Vector Machines (SVM), Decision Trees, and Naive Bayes have shown initial promise but cannot model contextual semantics, especially in longer or syntactically diverse text (Mukhtar and Khan, 2018). Deep learning models such as CNNs and LSTMs have improved contextual understanding, but they require large amounts of data to produce optimum outcomes (Ghulam et al., 2019; Chandio et al., 2022; Khan et al., 2022a).

One of the main issues in previous research is the lack of dedicated exploration of how LLMs can be harnessed to address the issues of Urdu sentiment analysis. While several studies have started to examine the application of pretext-trained transformers like BERT and XLM-R to Urdu sentiment classification (Ashraf et al., 2024, 2023; Khan et al., 2022b), less research has been conducted on how LLMs can be used not just for classification but also for data augmentation and cross-lingual transformation. Sentiment-bearing Urdu text collected from online sources is often naturally imbalanced,

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 198–207
March 28, 2026.

جذبہ: مشبت
(positive sentiment)
جملہ: یہ بہت خوشی کی بات
ہے کہ آج کا دن امید اور کامیابی
سے بھرپور ہے

جذبہ: منفی
(negative sentiment)
جملہ :یہ بات نہیں ہے ،
انکو پتہ چلا ہوں ، مجھے بھی
یقین نہیں ہے تو کل نہیں ہو

Figure 1: Example Urdu sentences illustrating positive and negative sentiment written in Urdu script. This figure highlights the linguistic structure and script characteristics of Urdu text commonly encountered in sentiment analysis tasks.

with a strong bias toward positive expressions. Furthermore, class imbalance is either disregarded or handled using traditional oversampling techniques rather than being managed with modern generative approaches. In addition, sentiment-bearing expressions in Urdu are often context-dependent and written in Nastaliq script, which differs substantially from Latin-based writing systems commonly used in high-resource languages. To illustrate the nature of Urdu text and the variation in sentiment expression, Fig. 1 presents example sentences written in Urdu script, representing positive and negative sentiments.

A further direction that has been less examined is where translation fits into sentiment analysis. English remains the language where most LLMs are trained and optimized. Because of this, English text benefits from more advanced tokenization techniques, more plentiful pretraining data, and generally superior model performance. This suggests a fascinating line of inquiry: would Urdu sentiment analysis performance be improved by first translating Urdu text to English and then applying high-performance English sentiment classifiers? The implications of this would be significant, especially for practical applications where performance and accuracy are of the utmost importance (Mukhtar and Khan, 2020; Saeed et al., 2024).

In this study, we sought to explore these two significant lines. First, we explore the impact of alleviating class imbalance in an Urdu sentiment dataset using LLMs-based data augmentation. We balance the class distribution by generating negative samples and analyzed the impact on model performance. Second, we test whether the translation of Urdu text to English and performing sentiment analysis in the English realm yields more favorable results than direct Urdu classification. Both of these approaches leverage the capability of LLMs but expand it in fundamentally different ways: one within the target language and the other through cross-lingual transformation.

We frame our study around two primary research questions; RQ1: What is the effect of LLM-based data augmentation on addressing class imbalance in Urdu sentiment classification? Here, we are interested in the effectiveness of synthetic data produced with LLM to augment the underrepresented class (typically negative sentiment). We evaluate whether supplementation with such data improves recall and F1-score for the minority class, thereby leading to more generalizable and balanced models. RQ2: How does translating Urdu text into English influence sentiment classification using English-specific LLMs?

To address these research questions, we conduct a series of controlled experiments on an Urdu sentiment dataset. We begin by setting an initial baseline performance on the original imbalanced dataset. We then introduce LLM-generated synthetic samples to mitigate class imbalance and reevaluate the models. Finally, we translate the dataset into English and apply English-specific sentiment classifiers to estimate gains. We employ accuracy, precision, recall, F1-score, and confusion matrix visualizations as metrics for detailed analysis. In summary, this study addresses key limitations of prior Urdu sentiment analysis research by empirically evaluating LLM-based augmentation and translation-based sentiment classification. By comparing data augmentation and translation-based techniques, we aim to provide insight into how sentiment analysis for low-resource languages can be significantly improved with the aid of current natural language processing advances.

2 Related Work

The sentiment analysis area for Urdu and Roman Urdu has witnessed significant growth from rule-based systems to high-end transformer-based models. However, the evolutionary process for these languages has not been in parallel with the developments in high-resource languages like English, and this is primarily due to the unique linguistic challenges and resource limitations presented by the Urdu language. These challenges include script differences, rich morphological inflections, and the absence of high-quality annotated datasets, which have collectively presented challenges to constructing strong and generalizable sentiment classification models (Khattak et al., 2021).

Earlier approaches have extensively employed traditional machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and k-nearest Neighbors (k-NN). These approaches relied on hand-crafted feature extraction methods such as TF-IDF, Bag-of-Words (BoW), and n-gram models to encode text data (Mukhtar and Khan, 2018). While these methods furnished baseline knowledge, their dependence on hand-crafted features also made them less generalizable to varied types of text and complex syntactic structures. The rigidity of these models typically made them less effective at handling the dynamics of informal language and domain-specific variations prevalent in social media and user-generated text (Farooq et al., 2023; Mukhtar and Khan, 2020).

The introduction of deep learning revolutionized Urdu sentiment analysis studies with deep models like Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and their bidirectional counterparts taking center stage (Chandio et al., 2022; Ghulam et al., 2019; Khan et al., 2021). These architectures facilitated automatic feature learning and improved the capability to model sequential dependencies in text. Studies that employed attention mechanisms brought the performance of models to new levels by allowing models to selectively focus on the most important parts of the input sequence (Naqvi et al., 2021; Khan et al., 2022a). Despite these advances, the performance of deep learning models has still been restricted by the lack of large-scale annotated Urdu datasets (Saeed et al., 2024).

More recently, transformer-based models have brought about a new revolution in the domain. Researchers have employed models such as mBERT, XLM-RoBERTa, and GPT-2 for Urdu sentiment analysis, yielding promising results (Ashraf et al., 2023, 2024; Khan et al., 2022b). These models leverage cross-lingual transfer learning through pretraining on multilingual corpora, which is the reason they achieve higher performance even in low-resource languages. XLM-R, for instance, has been applied to Roman Urdu sentiment classification by researchers and has been found to outperform recurrent and classical models (Ashraf et al., 2024; Khan et al., 2022a). Fine-tuning BERT and its multilingual variants has also been found to greatly enhance both binary and multi-class sentiment classification performance (Ashraf et al., 2023; Khan et al., 2022b).

Current research also explored prompt engineering and zero-shot/few-shot learning for sentiment analysis using large language models (LLMs). These approaches eliminate large volumes of labeled data and allow models to perform tasks via conditioning on carefully crafted prompts (Ahmed et al., 2024; Hasan et al., 2024). While this approach has been highly popular in English, it remains largely limited in Urdu. There is a strong opportunity to assess how LLMs respond to Urdu prompts and whether prompt-based learning can bridge the resource disparity in sentiment classification (Tahir et al., 2025).

One other prominent research area has been the creation and use of new datasets. Several studies have introduced custom datasets for Urdu and Roman Urdu sentiment analysis, usually collected from social media or review websites (Khan et al., 2022b; Shabbir and Majid, 2024). These datasets vary in terms of size and label granularity, with some of them being multi-class classification-friendly. Despite these efforts, the absence of standardized benchmarks continues to be a hindrance to advancing the field since it is challenging to compare results across studies or reproduce any results consistently (Khattak et al., 2021; Khan et al., 2022b).

Hybrid models that combine traditional and deep learning methodologies have also been explored. Some studies, for example, have combined CNNs with LSTMs or have used ensemble methods with decision trees and extra trees classifiers (Khan et al., 2022a; Ali et al., 2025; Saeed et al., 2024). These models attempt to combine the strengths of various algorithms to enhance predictive capability. While potent in some cases, they also involve greater computational complexity and require careful tuning.

Cross-lingual sentiment analysis efforts have also gained traction, spearheaded particularly by translation-based approaches. Here, Urdu texts are translated into English to leverage the strongly trained English sentiment classifiers (Mukhtar and Khan, 2020; Majeed et al., 2024; Saeed et al., 2024). This line of work has met with mixed success while translation introduces novel noise, in a few instances the better quality of English models can overshadow these limitations. Few studies have attempted to systematically quantify this trade-off, indicating an under-explored area with ample opportunity for improvement in Urdu sentiment analysis (Hasan et al., 2024).

Briefly, while the field has seen considerable progress with the introduction of deep learning and transformer-based approaches, several gaps remain. These include the absence of widespread application of LLMs to Urdu-specific tasks, limited exploration of prompt-based learning, inadequate addressing of class imbalance, and absence of benchmarking across studies. The current study addresses some of these gaps through an exploration of both LLM-based data augmentation and translation-based sentiment classification.

3 Methodology

We initiated by acquiring an Urdu sentiment dataset from an open-source repository (refer: (Batra et al., 2021)). The dataset was originally uploaded on Mendeley, a data-sharing platform, and needed strict preprocessing to make it conducive to sentiment classification. The raw data consisted of user-generated text in Urdu along with respective sentiment labels. The dataset was unstandardized and needed extensive cleaning.

3.1 Data Cleaning and Preprocessing

The text inputs contained some non-linguistic artifacts such as emojis, HTML entities, and aberrant punctuation, which were cleaned for consistency on input. More importantly, the sentiment labels were not provided as categorical classes. Instead, each example was labeled with a list of emotions or sentiment descriptors that had to be manually classified. Each descriptor was manually tagged with a binary tag of either "positive" or "negative" based on its semantic polarity and conventional sentiment interpretation. The mapping was performed manually at first and subsequently automated using a Python script.

Other preprocessing included the use of regular expressions to remove residual characters, emojis, and undesirable characters. We also normalized Urdu script characters for consistency and compatibility with tokenization. The dataset was cleaned and then checked to ensure that each entry was a well-formed text with a single binary sentiment label. The dataset was divided into training, validation, and test sets using an 80/10/10 split ratio. Stratified sampling was applied during splitting to preserve the original class distribution in all subsets. The training set was used to fine-tune the models, while the validation set was employed for early stopping and hyperparameter tuning. The test set, comprising 10% of the total data, was used for final evaluation of all models.

3.2 Tokenization and Model Architecture

To perform sentiment analysis, we employed a multilingual BERT model, specifically the `bert-base-multilingual-cased` model. The tokenizer is trained on a variety of languages, including Urdu, and processes complex scripts through WordPiece tokenization (Ashraf et al., 2023). WordPiece tokenizes text into subword pieces according to frequency in the training data, which allows it to handle out-of-vocabulary words and morphologically rich languages in a better way.

Tokenized text was used to fine-tune the multilingual BERT model on the binary sentiment classification task. It was trained using the Hugging Face Trainer API with evaluation metrics including accuracy, precision, recall, and F1-score, with primary emphasis on class-wise F1-score due to class imbalance.

3.3 Handling Class Imbalance Using LLMs

The initial dataset was highly imbalanced in terms of classes, with positive samples significantly outnumbering negative samples. To address this imbalance, we employed `bigscience/bloomz-1b1`, a multilingual instruction-tuned LLM, for data augmentation. Specifically, after performing a stratified 80/10/10 split, the training set contained 12,800 positive and 3,200 negative samples. We therefore generated 9,600 synthetic negative sentiment sentences in Urdu using carefully designed negative-sentiment prompts (Hasan et al., 2024).

All generated samples were passed through an aggressive cleaning pipeline to retain only structurally valid Urdu sentences, remove non-Urdu scripts, and enforce a minimum length requirement. After filtering, the augmented training set consisted of 12,800 positive and 12,800 negative samples, resulting in a balanced training corpus. Augmentation was applied exclusively to the training data, while the validation and test splits were kept unchanged, preserving their original class distributions. The multilingual BERT model was then further fine-tuned on this augmented training set and evaluated on the original stratified test set (see Fig. 2).

3.4 Translation and English LLM Evaluation

In order to explore the benefits of cross-lingual transfer, we translated the entire Urdu dataset into English via the `Helsinki-NLP/opus-mt-ur-en`

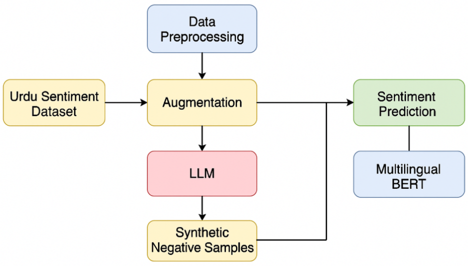

Figure 2: Pipeline for RQ1: Addressing Class Imbalance in Urdu Sentiment Analysis using LLM-Based Augmentation

translation model. The translation enabled us to leverage the English sentiment classification models that have high performance. We employed the `siebert/sentiment-roberta-large-english` model, a RoBERTa-based classifier that was fine-tuned for English sentiment tasks (Liu et al., 2019; Saeed et al., 2024).

We employed the same performance measures that were used for the Urdu models to assess the performance of the English pipeline. The English classifiers improved on a range of metrics, due to more advanced tokenization techniques and larger pretraining datasets available in English.

This method offers a step-by-step data preparation process, training models, class imbalance resolution through LLM-based augmentation, and cross-lingual translation for sentiment classification. By comparing performance on a baseline Urdu model, LLM-augmented Urdu model, and translated English pipeline, we aim to determine the most effective way of doing sentiment analysis under low-resource settings. In the following section, we present detailed experimental results and interpretation of our findings.

4 Experiments and Results

The experimental setup and evaluation results intended to answer the previously stated research questions are presented in this section.

4.1 RQ1: Impact of Class Imbalance Mitigation via LLMs

This experiment investigates the role of class imbalance in Urdu sentiment classification and evaluates whether large language models (LLMs) can be used to address it effectively. The original dataset we acquired was significantly skewed toward positive sentiments, with a ratio of 4:1 in favor of positive samples. This imbalance can cause classification models to perform disproportionately well on the majority class, while severely underperforming on the minority class.

4.1.1 Baseline Performance on Imbalanced Dataset

We first trained a multilingual BERT model (`bert-base-multilingual-cased`) (Ashraf et al., 2023) on the original imbalanced dataset. The training and validation splits maintained the original distribution, and the test set was likewise imbalanced. Table 1 presents the precision, recall, and F1-score for each class.

Table 1: Performance on Imbalanced Dataset (Baseline Model)

Class	Precision	Recall	F1-score
Negative	0.62	0.22	0.32
Positive	0.83	0.97	0.89

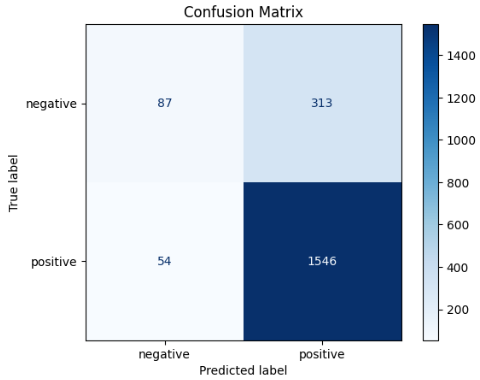

Figure 3: Confusion Matrix: Baseline (Imbalanced)

The results in Table 1 and Fig. 3 demonstrate that although the model achieves a high overall accuracy, it performs extremely poorly on the minority class. The negative class has a recall of only 0.22, which means that most negative samples are misclassified as positive. This indicates the model has learned to rely heavily on the dominant class and lacks the representational ability to generalize to minority class instances.

4.1.2 LLM-Based Data Augmentation

To mitigate class imbalance, we generated 9,600 synthetic negative sentiment samples using `bigscience/bloomz-1b1` (Hasan et al., 2024), a multilingual LLM instruction-tuned for generative tasks. Urdu prompts representative of negative sentiment were crafted to generate realistic examples. All generated sentences were then passed through a post-processing pipeline involving regular expression filtering, Urdu-script validation, and a minimum character threshold. These samples were combined with the 3,200 real negative samples present in the training split, resulting in a balanced training set of 12,800 negative and 12,800 positive examples.

We also experimented with other instruction-tuned language models, including Falcon-1 and AraGPT-2, as exploratory comparisons to contextualize the performance of our primary model. The latter has been developed specifically for Arabic text generation and shows promising capabilities in morphologically rich languages (Antoun et al., 2021).

The BERT model was retrained on this augmented dataset. Table 2 and Fig. 4 presents the classification metrics and confusion matrix on the test set.

Table 2: Performance on LLM-Augmented Balanced Dataset

Class	Precision	Recall	F1-score
Negative	0.87	0.79	0.83
Positive	0.93	0.96	0.95

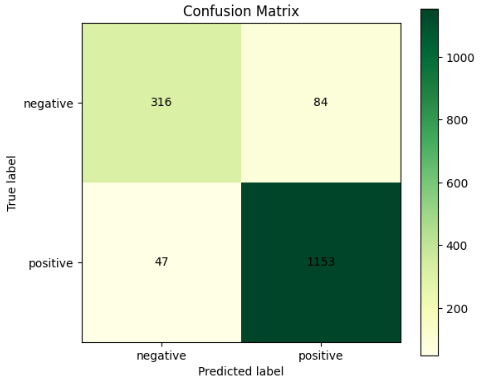

Figure 4: Confusion Matrix: LLM-Augmented Dataset

The LLM-augmented model exhibits a dramatic improvement in class-wise performance. Notably, the recall for the negative class improved from 0.22 (baseline) to 0.79, and the F1-score jumped from 0.32 to 0.83. These metrics confirm that the model, after augmentation, is significantly more effective in identifying the minority class, while maintaining high performance on the positive class. Table 3 reports the overall accuracy of the baseline and LLM-augmented models on the imbalanced test set.

Table 3: Accuracy Comparison for Class Imbalance Mitigation (RQ1)

Training Strategy	Acc
Imbalanced Data	0.82
LLM-Augmented	**0.92**

The analysis across two evaluation settings, imbalanced dataset and LLM-augmented balanced training set. It clearly reveals the transformative role of LLM-based data augmentation. While the baseline model leaned heavily on the dominant class, the augmented model displays well-balanced precision, recall, and F1-scores. This affirms that using LLMs to synthetically augment low-resource sentiment classes can be an effective and scalable solution to dataset imbalance. See Table 4 for negative-class performance across instruction-tuned LLMs.

Table 4: Performance Comparison of Class Imbalance Mitigation Models (on Negative Class)

Model / Strategy	Prec	Rec	F1	Acc
Baseline Imbalanced	0.62	0.22	0.32	0.82
`bloomz-1b1` (Final)	**0.87**	**0.79**	**0.86**	**0.92**
`falcon-rw-1b`	0.88	0.68	0.76	0.82
`aragpt2-base`	0.85	0.65	0.73	0.79

4.2 RQ2: Effect of Translating Urdu Sentences to English for Sentiment Classification

This research question investigates whether performance in Urdu sentiment classification can be improved by translating the text into English and using high-performing English sentiment classifiers. The rationale is grounded in the understanding that most large language models (LLMs), particularly transformer-based models, are primarily pretrained on English corpora, benefiting from richer linguistic features, superior tokenization strategies, and larger pretraining datasets.

4.2.1 Method Overview

To evaluate this hypothesis, we translated the entire Urdu dataset into English using the `Helsinki-NLP/opus-mt-ur-en` translation model. The translated dataset was then fed into siebert/sentiment-roberta-large-english, a RoBERTa-based classifier fine-tuned on English sentiment data (Liu et al., 2019). The classifier was not retrained or fine-tuned further and it was used in its zero-shot or direct inference capacity.

The evaluation was conducted on the test set consisting of 1600 examples, identical in content (but translated) to the test set used in RQ1.

4.2.2 Model Performance

The performance results are presented in Table 5. As shown, the classifier performed robustly on both sentiment classes, with closely aligned precision and recall values. Fig. 5 shows the confusion matrix for the translated dataset evaluated using the English LLM.

Table 5: Performance After Urdu-to-English Translation Using English LLM

Class	Precision	Recall	F1-score
Negative	0.81	0.78	0.79
Positive	0.92	0.93	0.92

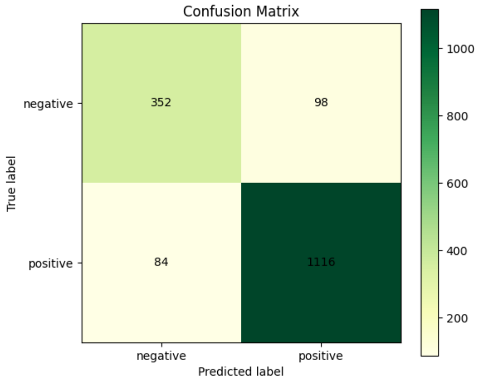

Figure 5: Confusion Matrix: Translated Dataset (English LLM)

4.2.3 Analysis and Comparison with Augmented Urdu Model

The experimental results provide insight into the effectiveness of both LLM-based data augmentation and translation-based cross-lingual sentiment classification for addressing challenges in Urdu sentiment analysis. When comparing the two approaches, distinct strengths and trade-offs emerge with respect to minority-class sensitivity and overall predictive performance.

The LLM-augmented Urdu model demonstrates strong gains in handling class imbalance. In particular, the negative class achieves a precision of 0.87, recall of 0.79, and an F1-score of 0.83, indicating a substantial improvement in minority-class recognition compared to imbalanced training scenarios. At the same time, the model maintains robust performance on the positive class (F1-score of 0.95), resulting in an overall accuracy of 0.92. These results suggest that synthetic data generation using instruction-tuned LLMs can effectively enhance the model's ability to learn discriminative features for underrepresented sentiment categories without sacrificing majority-class performance.

The translation-based approach, which leverages an English sentiment classifier after Urdu-to-English translation, exhibits competitive but slightly lower performance. While the positive class remains strong (F1-score of 0.92), the negative class shows comparatively reduced precision and recall (F1-score of 0.79), leading to an overall accuracy of 0.89. This performance gap suggests that although English LLMs benefit from richer pretraining and optimized tokenization, translation artifacts and potential semantic drift can still impact minority-class sentiment detection.

Overall, the findings indicate that LLM-based augmentation is particularly effective for mitigating class imbalance within the original language, while translation-based sentiment classification offers a viable alternative when high-resource language models are desired.

5 Discussion and Future Work

The experimental results highlight the complementary strengths of LLM-based data augmentation and translation-based cross-lingual sentiment classification for Urdu. Addressing RQ1, the LLM-augmented Urdu model demonstrates a clear improvement in handling class imbalance, particularly for the negative (minority) sentiment class. The observed gains in recall and F1-score indicate that instruction-tuned LLMs can effectively generate informative synthetic samples that help the classifier learn minority-class patterns more robustly. Importantly, these improvements are achieved without degrading performance on the positive class, suggesting that augmentation enhances class sensitivity rather than introducing bias.

The translation-based approach shows that mapping Urdu text into a high-resource language such as English can yield competitive sentiment classification performance. The strong results obtained using an English sentiment classifier reflect the advantages of richer pretraining data and more optimized tokenization available in English-centric models.Together, these results indicate that while cross-lingual transfer is a viable alternative, LLM-based augmentation within the source language may be more effective when minority-class sensitivity is a primary concern.

Looking ahead, several promising research directions emerge from this work. One avenue is to investigate the few-shot and zero-shot capabilities of newer instruction-tuned multilingual LLMs, such as BLOOMZ-MT and GPT-style models, for Urdu-specific sentiment analysis. Prompt-based evaluation in Urdu could reduce reliance on large labeled corpora or extensive synthetic data generation, offering a more lightweight alternative for low-resource scenarios (Ahmed et al., 2024; Hasan et al., 2024). Another important direction involves deeper analysis of translation-based sentiment classification. While our findings demonstrate the effectiveness of Urdu-to-English translation, further investigation is needed to understand how translation quality interacts with sentiment preservation. Techniques such as back-translation or selective manual annotation could help uncover systematic artifacts introduced during translation and their impact on downstream classification performance (Majeed et al., 2024). Additionally, extending this work beyond binary sentiment classification to include neutral, mixed, or fine-grained emotional categories would provide a richer understanding of sentiment expression in Urdu. Such extensions would require carefully defined labels and the development or adaptation of Urdu-specific datasets supporting multi-class annotations (Khattak et al., 2021).

6 Conclusions

This work explores the problem and potential solutions of sentiment analysis for Urdu, a language with very few annotated resources and class imbalance problems. Sensing the limitations of classical methods under low-resource conditions, we explore two complementing strategies that build on the strengths of state-of-the-art large language models (LLMs). The first method addresses class imbalance by synthetically creating negative samples from an instruction-adapted LLM. Such augmentation enables the creation of a more balanced and representative training set that can better allow a multilingual BERT classifier to learn minority sentiment patterns. The second method capitalizes on English language model maturity by translating the Urdu text into English and utilizing a pre-existing RoBERTa classifier for sentiment analysis.

Our comparative study of these approaches reveals that both approaches carry critical enhancements, but for varying aspects of the problem. The augmentation-based approach enhances sensitivity in the model towards underclass categories, and the translation-based approach leverages the robustness of high-resource models. The combined effect of these solutions provides us with an applied blueprint for sentiment analysis in linguistically underclass contexts. In addition to Urdu, this work's insights extend more generally to multilingual NLP activities.

7 Limitations

Despite the encouraging results, several limitations of this study should be acknowledged. First, while translation-based sentiment classification proved effective, the extent to which semantic fidelity is preserved during Urdu-to-English translation remains uncertain. Translation artifacts may introduce subtle distortions that affect sentiment interpretation and classification performance (Majeed et al., 2024). Second, the study focuses on binary sentiment classification, which may not fully capture the nuanced emotional expressions present in natural Urdu text. Extending the analysis to multi-class or fine-grained sentiment categories would require carefully defined labels and additional Urdu-specific annotated resources (Khattak et al., 2021).

Furthermore, although LLM-based augmentation mitigates class imbalance, the synthetic data generated may differ in subtle ways from organically produced language. Evaluating such data using qualitative linguistic analysis and human judgment, in addition to quantitative metrics, could provide deeper insight into its fluency and contextual appropriateness. Finally, accuracy may not fully reflect model behavior under class imbalance, which is why class-wise metrics are emphasized throughout the analysis. Moreover, the robustness of the proposed models to noisy, informal, and user-generated Urdu text characterized by spelling variation, transliteration, and non-standard grammar remains an open challenge.

Acknowledgments

The authors would like to thank Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum & Minerals (KFUPM) for supporting this work through SDAIA-KFUPM Joint Research Center for Artificial Intelligence grant number JRC-AI-CAI02563.

References

Rabbia Ahmed, Sadaf Abdul Rauf, and Seemab Latif. 2024. Leveraging large language models and prompt settings for context-aware financial sentiment analysis. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–9. IEEE.

Abid Ali, Mehmood Ul Hassan, Muhammad Munwar Iqbal, and Habib Akbar. 2025. Harnessing supervised machine learning for sentiment analysis in urdu text. In *AI-Driven: Social Media Analytics and Cybersecurity*, pages 45–60. Springer.

Amna Altaf, Muhammad Waqas Anwar, Muhammad Hasan Jamal, Sana Hassan, Usama Ijaz Bajwa, Gyu Sang Choi, and Imran Ashraf. 2022. Deep learning based cross domain sentiment classification for urdu language. *IEEE Access*, 10:102135–102147.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Aragpt2: Pre-trained transformer for arabic language generation. *Preprint*, arXiv:2012.15520.

M. R. Ashraf, M. Hussain, M. A. Jaffar, W. Y. Ramay, and M. Faheem. 2024. Revolutionizing urdu sentiment analysis: Harnessing the power of xlm-r and gpt-2. *IEEE Access*, 12:99779–99793.

M. R. Ashraf, Y. Jana, Q. Umer, M. A. Jaffar, S. Chung, and W. Y. Ramay. 2023. Bert-based sentiment analysis for low-resourced languages: A case study of urdu language. *IEEE Access*, 11:110245–110259.

Rakhi Batra, Zenun Kastrati, Ali Shariq Imran, Sher Muhammad Daudpota, and Abdul Ghafoor. 2021. A large-scale tweet dataset for urdu text sentiment analysis.

B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber. 2022. Attention-based ru-bilstm sentiment analysis model for roman urdu. *Applied Sciences*, 12(7):3641.

Muhammad Shoaib Farooq, Ansar Naseem, Furqan Rustam, and Imran Ashraf. 2023. Fake news detection in urdu language using machine learning. *PeerJ Computer Science*, 9:e1353.

Hussain Ghulam, Feng Zeng, Wenjia Li, and Yutong Xiao. 2019. Deep learning-based sentiment analysis for roman urdu text. *Procedia computer science*, 147:131–135.

Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *Preprint*, arXiv:2408.02237.

Lal Khan, Ammar Amjad, Kanwar Muhammad Afaq, and Hsien-Tsung Chang. 2022a. Deep sentiment analysis using cnn-lstm architecture of english and roman urdu text shared in social media. *Applied Sciences*, 12(5):2694.

Lal Khan, Ammar Amjad, Noman Ashraf, and Hsien-Tsung Chang. 2022b. Multi-class sentiment analysis of urdu text using multilingual bert. *Scientific Reports*, 12(1):5436.

Lal Khan, Ammar Amjad, Noman Ashraf, Hsien-Tsung Chang, and Alexander Gelbukh. 2021. Urdu sentiment analysis with deep learning methods. *IEEE access*, 9:97803–97812.

A. Khattak, M. Z. Asghar, A. Saeed, I. A. Hameed, S. A. Hassan, and S. Ahmad. 2021. A survey on sentiment analysis in urdu: A resource-poor language. *Egyptian Informatics Journal*, 22(1):53–74.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Adil Majeed, Usama Imtiaz, M. Asif Nseem, Muhammad Aleem, Waseem Shahzad, Mirza Omer Beg, and Hasan Mujtaba. 2024. Extracting emotion from resource poor language through transfer learning. *Multimedia Tools and Applications*.

N. Mukhtar and M. A. Khan. 2018. Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02):1851001.

N. Mukhtar and M. A. Khan. 2020. Effective lexicon-based approach for urdu sentiment analysis. *Artificial Intelligence Review*, 53:2521–2548.

U. Naqvi, A. Majid, and S. A. Abbas. 2021. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access*, 9:114085–114094.

Muhammad Saeed, Naeem Ahmed, Danish Ali, Muhammad Ramzan, Muzamil Mohib, Kajol Bagga, Atif Ur Rahman, and Ikram Majeed Khan. 2024. In-depth urdu sentiment analysis through multilingual bert and supervised learning approaches. *IECE Transactions on Intelligent Systematics*, 1(3):161–175.

Mamoona Shabbir and Muhammad Majid. 2024. Sentiment analysis from urdu language-based text using deep learning techniques. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–5. IEEE.

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking the performance of pre-trained LLMs across Urdu NLP tasks. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 17–34, Abu Dhabi, UAE. International Committee on Computational Linguistics.