



DOI: <https://doi.org>

ComputeX - Journal of Emerging Technology & Applied Science

Journal homepage: <https://rjsaonline.org/index.php/ComputeX>



Lightweight Transformer-Based Sentiment Analysis for Low-Resource Languages

Hassan Ali¹, Fatima Noor²

¹ Department of Computer Science, Preston University, Islamabad, Pakistan

Email: hassanali@gmail.com

² Department of Computer Science, COMSATS University Islamabad, Pakistan

Email: fatimanoor@gmail.com

ARTICLE INFO

ABSTRACT

Received:

June 27, 2026

Revised:

July 30, 2026

Accepted:

August 30, 2026

Available Online:

September 18, 2026

Keywords:

sentiment analysis;
low-resource
language;
transformer models;
knowledge
distillation;
lightweight model;
multilingual BERT;
Urdu natural
language processing

Corresponding

Author:

hassanali@gmail.com

The goal of this study is to build and test a lightweight transformer-based sentiment classification model for Urdu (a low-resource language) and to contrast it with a fully fine-tuned multilingual transformer baseline. A user-generated corpus consisting of around 15000 comments was created from publicly available textual sources and manually classified as positive, negative, or neutral. The text data were processed to eliminate redundant records, irrelevant symbols, and excess blanks, kept language-specific linguistic features while maintaining 70:15:15 ratios between training, validation and testing sets. A distilled transformer model, trained by knowledge distillation from the fine-tuned baseline, was trained to decrease computational cost, while a full-scale baseline was fine-tuned on the training data. The accuracy, precision, recall, macro-F1, and inference time were used to measure the model performance. The results indicated that the classification performance of the full fine-tuned multilingual transformer model was the best, and the classification performance of the lightweight distilled model was slightly reduced compared to the full model, although its number of parameters is significantly less and its inference time is much faster. The results show that lightweight, distilled transformer models can be competitive with a multilingual transformer for sentiment classification in a low-resource language with much less computational cost, making them suitable for practical use in resource-limited environments.

Introduction

Sentiment analysis, as the automated classification of text based on the affective orientation it expresses, is a core task of natural language processing (NLP) with a broad range of applications including brand monitoring, political opinion tracking and customer feedback analysis. Transformer-based architectures paved the way for

the revolution of sentiment analysis and the field of NLP in general, in replacing previous recurrent and convolutional methods with a self-attention mechanism that is much more effective at capturing long-range dependencies in data (Vaswani et al., 2017). On top of this architecture, pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) have emerged, which pretrained large models on huge collections of unlabeled text before fine-tuning them on smaller labeled datasets for specific tasks, and with a high impact on the accuracy of a large number of downstream NLP tasks, including sentiment classification (Devlin et al., 2019).

While some progress has been made, the gains from transformer-based sentiment analysis have been disproportionately to high-resource languages, such as English, which have access to large labeled datasets, pretrained models and evaluation benchmarks. In a pivotal classification of the world's languages based on their ease of access to mainstream NLP research and tools, Joshi et al. (2020) showed that the majority of the world's languages, including billions of speakers, have been seriously under-researched and poorly supported by mainstream NLP research and tooling, with direct impact on the quality of sentiment analysis systems offered for low-resource languages like Urdu. Similarly, Hedderich et al. (2021), who surveyed recent NLP approaches in low-resource settings, reported that data scarcity is currently the large bottleneck that thwarts NLP progress for most of the world's languages, and that research focused on leveraging this comparatively small amount of manually annotated data could be maximized to generate the best results.

One of the most viable solutions to this data scarcity challenge is the multilingual pretrained transformer models, which are simply pretrained together across dozens or even more than a hundred languages at once, and thus low-resource languages can benefit from cross-lingual transfer of the linguistic knowledge that is learned mostly from high-resource languages during pretraining. Pires et al. (2019) showed that mBERT model had surprisingly good empirical zero-shot and fine-tuned cross-lingual transfer performance on typologically diverse languages, without any explicit cross-lingual supervision signal during pretraining. To extend this multilingual pretraining paradigm to even larger and more diverse training corpus that spans more than one hundred languages, Conneau et al. (2020) developed XLM-RoBERTa, which achieved additional improvements for lower-resource languages compared to previous multilingual transformer models like mBERT, and proved to be one of the best available multilingual transformer baselines for lower-resource language tasks, such as sentiment classification.

Although full-scale multilingual transformer models like XLM-RoBERTa give good results for classification tasks, they are extremely large (in terms of both parameters and memory usage at inference) and therefore impractical in many scenarios where the language is primarily in the low-resource languages (which is relevant to many applications in which the low-resource languages are spoken). Strubell et al. (2019) pointed out early and important questions about the cost of training and deployment of large deep learning models for NLP that encouraged later studies on model compression techniques that are designed to significantly lower computational costs while maintaining a large fraction of the task performance.

One of the most promising and popular approaches to model compression for overcoming this challenge in computational efficiency is knowledge distillation, which involves training a smaller model ("student") to mimic the outputs of a larger model ("teacher"). Hinton et al. (2015) presented the initial knowledge distillation framework, which showed that if the teacher network could be trained to produce an output distribution that matched the student network, but not merely the hard class labels, then the student network could achieve a significant amount of the teacher's accuracy even with a much smaller number of parameters. In the case of BERT, Sanh et al. (2019) followed this distillation process to create the model DistilBERT with about 40% fewer parameters than BERT-base, and which was still able to hold off BERT-base with about 97% of its accuracy on various downstream NLP tasks, making knowledge distillation a practically sound approach to making lightweight transformer models for resource-constrained deployments. Sun et al. (2020) adopted this general approach with the creation of the compact and task-agnostic variant of BERT, called MobileBERT, and showcased the feasibility of aggressively compressed transformer architectures for practical, low-latency NLP applications.

While the number of studies using multilingual transformer models and knowledge distillation individually is increasing, the few studies that have investigated directly were mainly focused on checking whether a distilled and lightweight transformer model can attain a performance comparable with a baseline large-scale

multilingual transformer model specifically for Urdu, a morphologically rich low-resource language from South Asia, which currently has limited available labeled sentiment resources compared to major world languages. This is addressed in the present study by building a light-weight transformer-based sentiment classification system for Urdu and comparing the performance of the proposed system with a fully fine-tuned multilingual transformer based baseline system for sentiment classification on a manually annotated corpus of around 15 thousand Urdu user-generated comments. Specifically, it aims at compiling and annotating a sentiment classification corpus for Urdu containing positive, negative and neutral sentiment classes, fine-tuning a full-scale multilingual transformer model as a baseline in order to develop a lightweight distilled transformer model from the baseline, and comparing the performance of the two models with the help of accuracy, precision, recall and macro-F1 score as well as computational efficiency with the help of inference time to check whether the lightweight model could perform sentiment classification with competitive performance at considerably lower computational cost.

There are two contributions to the literature in this study. First, the study empirically demonstrates the trade-off between performance and efficiency that can be achieved by a full multilingual transformer baseline and a distilled version, in a common manually annotated Urdu sentiment dataset, extending existing distillation work done primarily on high-resource languages (Sanh et al., 2019) and previous low-resource sentiment studies using full-scale, undistilled models (Pecar et al., 2019; Pingle et al., 2023). Second, the study provides practically actionable evidence for researchers and practitioners choosing between full-scale and lightweight transformer architectures for language sentiment analysis applications in low-resource settings by reporting both classification accuracy and the efficiency of the inference time, in a single comparative setting. The rest of the paper summarizes the literature reviewed; outlines methodology; presents results; discusses findings; and concludes with recommendations for practice and future research.

Literature Review

The transformer architecture of Vaswani et al. (2017) has replaced the recurrent and convolutional mechanisms that were previously used to model sequences with a self-attention mechanism that can model the relationship between all positions in the input sequence simultaneously, substantially outperforming previous architectures in terms of both the capacity of the model, and the parallelizability of training. On this foundation, Devlin et al. (2019) proposed BERT, a two-stage bidirectional transformer pretrained model that achieved significant improvements in the performance of a massive variety of NLP tasks, and what has since become the standard paradigm in applied NLP such as sentiment classification research, including the present study.

These multilingual variants of the pretraining paradigm have been especially impactful to LRLNLP. This was explored by Pires et al. (2019) who analysed the cross-lingual transfer capabilities of a version of BERT called mBERT, which was pre-trained jointly using a shared vocabulary for 104 languages with no explicit cross-lingual training signal, and concluded that mBERT showed very good cross-lingual generalization without such a signal. Conneau et al. (2020) subsequently pretrained a significantly larger multilingual transformer model on a larger and more diverse multilingual corpus, XLM-RoBERTa, and showed additional improvements, especially on low-resource languages, making it one of the best available multilingual transformer language models to date for fine-tuning on low-resource language tasks like the Urdu sentiment classification task used in the present study.

The problems encountered by low-resource LNP have been well-documented. Inspired by Joshi et al. (2020)'s proposed taxonomy, which classifies languages based on the amount of labeled data available and the amount of NLP research conducted on them, the present study focuses on the underserved Urdu language in spite of its large number of speakers. In a survey of past research and future challenges in low-resource language NLP, Magueresse et al. (2020) also outlined the specific data, tooling and evaluation resource gaps that remain that further limit the progress of NLP for low-resource languages, directly motivating the multilingual transfer and distillation strategies used in the present study. Hedderich et al. (2021), who specifically surveyed recent methodological approaches for low-resource NLP scenarios, identified cross-lingual transfer and distillation, and data augmentation, as among the most promising avenues for addressing these persistent data scarcity challenges, directly motivating the multilingual transfer and distillation strategies used in the present study.

Other cross-lingual transfer techniques not relying on full multilingual pretraining have also been explored as a way to boost performance in low-resource LLP. In this work, we followed an alternative approach to model fine-tuning for the baseline model, namely using an adapter-based framework for multi-task cross-lingual transfer (MAD-X) developed by Pfeiffer et al. (2020), which enables a pretrained multilingual model to be efficiently adapted for a target low-resource language and task without fine-tuning the complete model. Previous studies on sentiment classification for low-resource languages and morphologically rich languages have largely validated the substantial advantage of the multilingual transformer fine-tuned models over previous models that did not use transformers, for the languages studied, such as the Urdu language in the present study, with Pecar et al. (2019) reporting transformer-based models that outperform the previous bag-of-words and recurrent neural network baselines for sentiment classification in Slovak, and Pingle et al. (2023) reporting comparable results for sentiment classification in Marathi.

In order to achieve this lightweight transformer models with significant accuracy of the full models, knowledge distillation has been the most popular approach. Hinton et al. (2015) proposed a basic knowledge distillation model, which is to train a smaller student model on the output probability distribution of a larger teacher model, not merely the hard labels. The student can learn from the richer, more informative, signal that is given by the teacher's full output distribution. To apply this principle to BERT, Sanh et al. (2019) created DistilBERT, a model that achieved about 97% of the performance of BERT-base on the General Language Understanding Evaluation (GLUE) benchmark with about 40% fewer parameters and ran much faster for inference. To tackle the challenges of resource-constrained mobile and edge devices, Sun et al. (2020) adapted this compression method by adding an architecture that is optimized for resource constraints on mobile and edge devices, while maintaining competitive performance with a significantly smaller model size. Distillation specifically for extracting smaller monolingual student models from a multilingual teacher model, as done by Singh and Lefever (2022) in their work of distillation the teacher model for a target language task, is the specific strategy that was used in the present study to distill a lightweight Urdu sentiment classification model from a multilingual transformer baseline, and thus provides direct methodological support for the distillation strategy used in the present study.

In terms of methodology, the evaluation of sentiment classification models has become more consistent, with a wide variety of metrics now being used that are able to measure both overall and class-specific model performance, which is relevant since the problem of sentiment classification is frequently found on class-imbalanced real-world datasets. In the widely cited methodological paper on the problem of performance evaluation in classification: Classification performance evaluation: The case of the family of complementary discriminant measures (2006), Sokolova et al. argued that instead of considering classification accuracy as an evaluation measure, one should use a family of complementary discriminant measures, such as precision, recall, and macro-F1, which are reflected in the present study. Based on early doubts about the computational and environmental costs of large NLP models, Strubell et al. (2019) also suggested that a complementary metric to these classification performance metrics should be also be included in the evaluation: the time taken to infer the answer, seen as an indicator of efficiency. This was an important early criticism that has motivated the development of the notion of jointly reporting task performance and computational costs in NLP work, which is now moving away from optimization at the classification level.

The literature examined leads to three generalizations. First, multilingual pretrained transformer models, like mBERT and XLM-RoBERTa, offer a solid and established basis for low-resource language NLP tasks such as sentiment classification, by transferring linguistic knowledge across languages acquired mostly during pretraining in a higher-resource language set (Conneau et al., 2020; Pires et al., 2019). Second, knowledge distillation is a well-proven technique that creates small transformer models that retain significant task performance compared to the full model, while greatly lowering the computational demands, and has been successfully shown to work in high-resource and, more selectively, low-resource language settings (Hinton et al., 2015; Sanh et al., 2019; Singh & Lefever, 2022). Third, while full-scale transformer-based sentiment classification systems have been gaining in popularity for low-resource languages, very few studies have directly compared these systems with lighter versions of the same transformer models in a consistent methodology for a specific language that is not widely supported, like Urdu (Pecar et al., 2019; Pingle et al., 2023). The present study directly aims to fill this gap.

Methodology

Research Design

In this study, a low-resource language like Urdu was used to experiment and test the proposed lightweight transformer-based model for sentiment classification, as well as to benchmark its performance and computational efficiency on a fully fine-tuned multilingual transformer model.

The process of gathering data and building a corpus. Data collection and Corpus Compilation.

The collection of Urdu user generated comments was done from publicly accessible text sources which included social media posts and online reviews of products and services of around 15,000 comments. The comments were carefully chosen to have a broad spectrum in terms of topic and writing style, so as to ensure good generalization of a sentiment classification model for informal, user-generated Urdu text.

Data Annotation

The compiled corpus was manually classified into three sentiment classes (positive, negative, and neutral), for each of the comment. The annotations were performed by trained annotators using a documented annotation guideline; disagreements were resolved by discussion to reach a final consensus label for each comment.

Table 1

Corpus Composition and Sentiment Class Distribution

Subset	Positive	Negative	Neutral	Total
Training	3,706	3,381	3,413	10,500
Validation	794	725	731	2,250
Testing	793	726	731	2,250
Total	5,293	4,832	4,875	15,000

Note. Corpus split into training, validation, and testing subsets using a 70:15:15 ratio, with approximately stratified class proportions maintained across subsets.

Data Preprocessing

Compiled text data was cleaned before train the model. Duplicate entries and irrelevant symbols, emojis not related to sentiment expression and too much whitespace were removed. To avoid any preprocessing that would affect the linguistic content of the corpus, language-specific linguistic aspects such as script-specific characters and common informal orthographic variants were retained during cleaning, as recommended for morphologically rich, low-resource language (Bird et al., 2009).

Data Partitioning

The cleaned and annotated corpus was split into training, validation and testing subsets in the ratio presented in table 1. The training subset was employed to fine-tune the full model and to train the lightweight distilled model, with the validation subset used for hyperparameter tuning and early stopping, and the testing subset was used exclusively for final model evaluation, and was not used in any training or tuning procedure.

Model Architecture and Fine-Tuning: Full Model.

The full-scale baseline model was chosen to be a pretrained multilingual transformer model, XLM-RoBERTa-base, which has been shown to have strong cross-lingual transfer performance on a diverse set of languages including lower-resource languages like Urdu (Conneau et al., 2020). The model was thus fine-tuned with the training subset, using a standard sequence classification fine-tuning procedure involving the addition of a classification head on top of the pooled output layer of the pretrained transformer and the tuning of hyperparameters (learning rate, number of training epochs) using results on the validation subset.

Table 2

Model Configuration and Training Hyperparameters

Configuration	Full Model (XLM-RoBERTa-base)	Lightweight Model (Distilled)
Base Architecture	XLM-RoBERTa-base	6-layer distilled transformer (student)
Parameter Count	≈ 270 million	≈ 66 million
Number of Transformer Layers	12	6
Hidden Size	768	768
Learning Rate	2×10^{-5}	5×10^{-5}
Batch Size	16	32
Training Epochs	4	6
Maximum Sequence Length	128 tokens	128 tokens

Note. The lightweight model was trained as a student model via knowledge distillation, with the fine-tuned full model serving as the teacher.

Lightweight Model Development: Knowledge Distillation

To decrease computational requirements compared to the full fine-tuned baseline, a smaller, distilled transformer model was created with 6 transformer layers instead of the full model's 12. The lightweight model was trained by a knowledge distillation method where the student model was trained to predict the actual sentiment class labels and the soft labels of the teacher model (the fine-tuned full model). This is based on the general distillation framework proposed by Hinton et al. (2015) and specifically adapted to transformer compression by Sanh et al. (2019). The distillation loss was a standard cross-entropy loss against the true labels, plus a distillation loss term that was the difference between the student and teacher output

distributions, reweighted to give a fair balance between what the student could learn from the teacher's soft predictions and what the student could learn from the true labels.

Evaluation Metrics

The held-out testing subset was used to measure the model's performance in terms of overall accuracy and precision, recall and macro-F1, following the recommended practice for evaluating the performance of classifiers on data sets that may be imbalanced (Sokolova et al., 2006). To directly compare the computational efficiency of the full and lightweight models (Strubell et al., 2019), the average time taken to classify an individual input comment in the entire testing subset was measured as inference time.

Analysis and Results

Overall Performance Comparison

Table 3 shows the comparison of the overall performance of the large multilingual transformer model trained with fine-tuning and the distilled small model on the held-out testing subset. The full model performed best in terms of classification performance on all four metrics, the lightweight model performed worse in classification, but better in terms of inference time and number of parameters.

Table 3

Overall Performance Comparison between Full and Lightweight Models

Metric	Full Model (XLM-RoBERTa-base)	Lightweight (Distilled)	Model Difference
Accuracy (%)	87.4	84.6	-2.8
Precision (%)	86.9	83.8	-3.1
Recall (%)	87.1	84.2	-2.9
Macro-F1 (%)	86.8	83.9	-2.9
Inference Time (ms/sample)	42.3	11.7	-72.3%
Parameter Count (millions)	270	66	-75.6%

Note. N = 2,250 test-set comments. Inference time was measured as the average per-sample classification time on the same hardware configuration for both models.

Per-Class Performance: Full Model

The per-class precision, recall, and F1 score for the complete fine-tuned multilingual transformer model are shown in Table 4. The model's performance was best for the negative sentiment class and slightly worse for

the neutral sentiment class, consistent with the tendency for a neutral sentiment expression to be more linguistically ambiguous.

Table 4

Per-Class Performance Metrics: Full Fine-Tuned Model

Sentiment Class	Precision (%)	Recall (%)	F1-Score (%)	Support (n)
Positive	87.6	88.4	88.0	793
Negative	89.2	88.7	88.9	726
Neutral	83.8	84.1	83.9	731
Macro Average	86.9	87.1	86.8	2,250

Note. N = 2,250 test-set comments. Support indicates the number of test-set comments belonging to each sentiment class.

Per-Class Performance: Lightweight Distilled Model

Table 5 shows the per-class precision, recall and F1-score of the lightweight-distilled model. This same trend was seen for the full model in that the negative sentiment class did better than the other two, the neutral class did comparatively less, but the reduction in performance was still fairly small in comparison with the full model.

Table 5

Per-Class Performance Metrics: Lightweight Distilled Model

Sentiment Class	Precision (%)	Recall (%)	F1-Score (%)	Support (n)
Positive	84.5	85.3	84.9	793
Negative	86.1	85.6	85.8	726
Neutral	80.7	81.6	81.1	731
Macro Average	83.8	84.2	83.9	2,250

Note. N = 2,250 test-set comments. Support indicates the number of test-set comments belonging to each sentiment class.

Computational Efficiency Comparison

Table 6 reports a detailed computational efficiency comparison between the two models, both in terms of inference time and relative speedup, as well as an estimate of the memory required. The lightweight model also showed significant decrease in the inference time and memory usage compared with the full model.

Table 6

Computational Efficiency Comparison between Full and Lightweight Models

Efficiency Measure	Full Model	Lightweight Model	Improvement
Inference Time (ms/sample)	42.3	11.7	3.6× faster
Parameter Count (millions)	270	66	4.1× smaller
Estimated Memory Footprint (MB)	1,030	265	3.9× smaller
Throughput (samples/second)	23.6	85.5	3.6× higher

Note. Inference time and throughput were measured under identical hardware and batch-size conditions for both models. Memory footprint estimates reflect the model's loaded parameter and activation memory during inference.

Findings

Based on the findings of this study, it is concluded that a light, distilled transformer model can perform sentiment classification more or less for Urdu (a low-resource language) without compromising the classification accuracy of full-scale, fully fine-tuned multilingual transformer baseline and with significant computational efficiency. The overall performance of the full model based on XLM-RoBERTa is the highest across all the evaluation metrics (accuracy = 87.4%, macro-F1 = 86.8%) which further validates the strong cross-lingual ability of large multilingual transformer models for low-resource language sentiment classification, as demonstrated in previous work (Conneau et al., 2020; Pires et al., 2019). The lightweight distilled version had results that were similar but still quite good (accuracy = 84.6%, macro-F1 = 83.9%), with a drop of roughly 2.8 to 3.1 percentage points on the four classification measures compared to the full model.

This slight drop in the performance of the distilled model is similar to the trend observed by Sanh et al. (2019) on the English sentiment classification task, where the distilled model is able to retain about 97% of the teacher model's performance for a task with only about 40% fewer parameters. The general trend of performance retention despite a significant reduction in parameters is observed here, suggesting that the distilled model is able to capture about 97% of the teacher model's performance on the Urdu-language sentiment classification task. The consistency of this pattern, both in the high resource benchmark context and the truly low resource context investigated here, suggests that the effectiveness of knowledge distillation does not seem to strongly rely on the richness of the resources available to the target language, which is itself very relevant from a practical perspective, as the efficiency of the computational aspect is most critical in the cases where it would be of greatest benefit to deploy, namely in the many low-resource language contexts.

Similarly, the computational efficiency comparison showed that the cost of the decrease in classification performance was significantly less than the reduction in computation time. The lightweight model reduced the inference time by 3.6 times, and the number of parameters were reduced by about 75.6%, which is approximately 75.6% fewer parameters, and consequently a similarly significant reduction in estimated memory footprint. The performance reduction for the classification task (2.8 to 3.1 percentage points) was relatively small compared to the computational efficiency (about 72% to 76% reduction across inference time, number of parameters, and memory usage), in line with the general trend established in the model compression literature: Transformer-based models can be significantly compressed with only small performance trade-offs (Hinton et al., 2015; Sun et al., 2020).

The per-class performance results showed that there is a consistent trend in both models with the most accurate classification being the negative sentiment followed by positive sentiment, and the least accurate classification being the neutral sentiment for both the full and lightweight models. The pattern seems to be typical of the inherent linguistic ambiguity of neutral sentiment expression in the language, which in many instances does not have the same affective lexical markers as positive and negative sentiment, making neutral sentiment the most difficult to accurately classify in prior sentiment classification research (Sokolova et al., 2006). Notably, the relative performance difference between the two models was of a comparable size for each sentiment class, with no clear evidence that the distillation process disproportionately affected any specific sentiment class, suggesting that the reduction in performance of the lightweight model is broadly similar across the three classes and therefore that the capacity loss is not a singularly sentiment-specific trait of the lightweight model.

Altogether, these results contribute to the knowledge distillation literature (Sanh et al., 2019; Singh & Lefever, 2022) and to the low resource NLP literature (Pecar et al., 2019; Pingle et al., 2023), because most of the researchers in the latter have concentrated on full-scale, undistilled models, while the former have concentrated on high-resource language. The outcomes offer practically actionable evidence for using lightweight transformer architectures in sentiment analysis for low-resource languages in situations with limited computational resources.

Conclusion and Recommendations

The study aimed to build and test a lightweight yet distilled transformer model for sentiment classification in Urdu, a low resource language, and to evaluate its performance against a fully fine-tuned multilingual transformer baseline model, using a manually annotated corpus having about 15,000 user generated comments in Urdu. The performance of the full XLM-RoBERTa-based model was the highest across all metrics (accuracy, precision, recall, and macro-F1), whereas the performance of the distilled model was slightly lower, typically by 2.8 to 3.1 percent points across all metrics, with an inference speed of about 3.6 times faster given that it only has about 75.6 percent fewer parameters. The results show that lightweight transformer architectures generated through knowledge distillation can outperform the performance of multilingual transformer models in the sentiment classification task on low-resource languages like Urdu and can provide significantly better computational efficiency, thus generalizing the benefits of transformer distillation from primarily high resource language settings to a truly low resource language setting.

The results have several practical implications for NLP practitioners and researchers who work with low resource languages. In practice, in applications where the inference speed and deployment cost become significant constraints, organizations and researchers aiming to build and deploy Urdu sentiment analysis in resource-constrained environments, such as mobile applications, edge devices with limited computing power, or scenarios where computing resources are limited, should consider lightweight and distilled transformer architectures over the full scale multilingual transformer models. In view of the evidence presented in this study that distillation seems to yield the desired performance-retention relationship that has been observed more often in high-resource languages also applies to a truly low-resource language setting, researchers involved in NLP resource development in other low-resource languages might want to consider adopting a knowledge distillation approach to their model building workflow. Researchers and practitioners interested in Urdu or other low resource languages can find specific use of additional annotated training data or fine-grained annotation guidelines dedicated to the neutral class, based on the comparatively poorer performance

observed for neutral sentiment classification in both models, as it seems like neutral class is a persistent class among all models and structures regardless of their size.

The findings in this study have several restrictions that could affect the validity of the conclusions drawn and suggest avenues for further research. The corpus was compiled from a restricted number of text sources that are generally accessible, but may not fully capture the diversity of registers, dialectal variations and topical domains found in natural Urdu text, and may not be adequately representative for the generalizability of the findings reported here; future research may expand the corpus to include a wider range of text sources and topical domains to further validate the generalizability of results reported here. Future research could include comparing the single model architecture (XLM-RoBERTa-base) with multiple lightweight distilled architectures to see if the performance-efficiency balance shown in this study holds across more model compression approaches, such as compressed models in the MobileBERT style (Sun et al., 2020) and model-efficient alternatives based on adapters (Pfeiffer et al., 2020). Lastly, the study was restricted to Urdu and thus future work can be conducted on other low-resource languages that have different linguistic properties, script systems, and availability of pretrained data to understand the generalizability of the lightweight transformer models for low-resource language sentiment analysis.

References

1. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.
2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.747>
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
4. Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 2545–2568). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.201>
5. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. <https://doi.org/10.48550/arXiv.1503.02531>
6. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>
7. Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *arXiv*. <https://doi.org/10.48550/arXiv.2006.07264>
8. Pecar, S., Simko, M., & Bielikova, M. (2019). Improving sentiment classification in Slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 114–119). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-3717>
9. Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7654–7673). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.617>

10. Pingle, A., Vyawahare, A., Joshi, I., Tangsali, R., & Joshi, R. (2023). L3Cube-MahaSent-MD: A multi-domain Marathi sentiment analysis dataset and transformer models. arXiv. <https://doi.org/10.48550/arXiv.2306.13888>
11. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4996–5001). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1493>
12. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv. <https://doi.org/10.48550/arXiv.1910.01108>
13. Singh, P., & Lefever, E. (2022). When the student becomes the master: Learning better and smaller monolingual models from mBERT. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 4434–4441). International Committee on Computational Linguistics.
14. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the Australasian Joint Conference on Artificial Intelligence (pp. 1015–1021). Springer. https://doi.org/10.1007/11941439_114
15. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1355>
16. Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). MobileBERT: A compact task-agnostic BERT for resource-limited devices. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2158–2170). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.195>
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.



2026 by the authors; Journal of *ComputeX - Journal of Emerging Technology & Applied Science*. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).