Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 1 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i01.58867

SHORT PAPER

Multi-Task Mining of Ethiopian Mobile App Reviews Using Machine Learning and Deep Learning Approaches

Alemu Kumilachew Tegegnie (✉)

Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia

alemu.kumilachew@bdu.edu.et

ABSTRACT

The rapid growth of mobile applications in Ethiopia has generated a wealth of user-generated content in the form of app reviews and ratings. These reviews provide critical insights into user satisfaction, app performance, and feature demands. However, systematic analysis of such unstructured and multilingual feedback remains limited in Ethiopia due to the absence of automated tools and localized natural language processing (NLP) resources. This study introduces a multi-task review mining framework that integrates sentiment classification, feedback categorization, and rating prediction. A dataset of 10,200 Ethiopian mobile app reviews collected from the Google Play Store was preprocessed, annotated, and analyzed using both machine learning and deep learning models. Experimental results indicate that convolutional neural networks (CNNs) outperformed other models, achieving 98.7% accuracy for sentiment classification, 96.6% for feedback categorization, and an R^2 of 0.40 for rating prediction. Among traditional models, XGBoost demonstrated strong performance, particularly in classification tasks. The findings highlight the effectiveness of CNN-based models in extracting actionable insights from multilingual reviews, offering developers and policymakers data-driven tools to improve app quality and enhance user satisfaction. This study contributes to the growing field of opinion mining in low-resource contexts and aligns with Ethiopia's Digital Transformation 2025 agenda.

KEYWORDS

mobile app reviews, sentiment analysis, feedback categorization, rating prediction, natural language processing (NLP), deep learning, Ethiopia

1 INTRODUCTION

The global expansion of mobile technologies has profoundly reshaped service delivery across multiple sectors [1]. As of 2023, consumer spending on mobile applications exceeded 170 billion U.S. dollars [2], reflecting their growing economic and social impact. Ethiopia is no exception to this trend. Through the Digital Ethiopia 2025 strategy, the Ethiopian Ministry of Innovation and Technology [3] has prioritized mobile applications as enablers of digital transformation in finance, healthcare, transportation,

Tegegnie, A. K. (2026). Multi-Task Mining of Ethiopian Mobile App Reviews Using Machine Learning and Deep Learning Approaches. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(1), pp. 160–167. https://doi.org/10.3991/ijim.v20i01.58867

Article submitted 2025-10-06. Revision uploaded 2025-11-25. Final acceptance 2025-11-25.

and commerce. These applications are increasingly relied upon by citizens for essential services such as mobile money transfers, ride-hailing, and online shopping. User reviews are a rich source of data for evaluating these apps. Prior studies demonstrate that reviews not only indicate user satisfaction but also provide developers with crucial feedback for bug fixing, performance improvement, and feature enhancement [4]. For instance, a study by [5] revealed that more than half of app reviews contain actionable information, such as complaints or feature requests that directly inform development priorities. In Ethiopia, however, the utility of such reviews remains largely untapped. Developers often manually browse through user feedback, a process that is inefficient and error-prone given the volume of reviews. Furthermore, linguistic complexities such as code-switching between Amharic and English exacerbate the difficulty of systematic analysis. This study addresses these challenges by introducing a multi-task review mining framework. The framework integrates three complementary tasks: sentiment classification, semantic feedback categorization, and rating prediction. The rationale for adopting a multi-task approach lies in its ability to provide a holistic understanding of user perspectives. Sentiment classification helps gauge overall satisfaction; feedback categorization identifies specific improvement areas; and rating prediction links textual sentiment with numerical app ratings, offering an additional quality metric. Together, these tasks provide developers and policymakers with actionable insights to strengthen Ethiopia's growing app ecosystem.

2 LITERATURE REVIEW

2.1 Opinion mining and sentiment analysis

Opinion mining, or sentiment analysis, has become an essential component of natural language processing (NLP). It enables the automatic identification of attitudes and opinions expressed in text. [6], [7] highlight that sentiment analysis supports industries ranging from marketing to e-governance by converting unstructured data into structured insights. Traditional machine learning techniques such as Naïve Bayes, support vector machines (SVM), and random forest classifiers have long dominated sentiment analysis, achieving moderate success with structured datasets. However, their reliance on handcrafted features often limits performance in multilingual and noisy contexts. The rise of deep learning has significantly improved text analysis. LSTM and Bi-LSTM models capture long-range dependencies in text, which is especially useful in sentiment detection [8]. Convolutional neural networks (CNNs), initially developed for image recognition, have also been successfully applied to textual data by capturing local n-gram features. They have proven particularly effective for short and informal texts, such as tweets or app reviews [9]. More recently, Transformer-based models such as BERT [10] have set new benchmarks in sentiment analysis. However, these models are resource-intensive and often impractical in low-resource contexts such as Ethiopia.

2.2 App review mining

App review mining specifically has gained scholarly attention due to its practical value. [4] show that review analysis helps developers triage bug reports and prioritize feature development. Similarly, [11] argue that reviews act as a form of crowdsourced quality assurance. Feedback categorization, in particular, has been found to enhance developers' responsiveness to user needs [5], [12], [13].

2.3 Ethiopian context

In Ethiopia, research remains at an early stage. [14] examined Amharic reviews using traditional ML methods, achieving only moderate classification accuracy. Their study was constrained by a small dataset, reflecting the scarcity of annotated resources in local languages. [15] demonstrated that deep learning models, particularly CNN and LSTM, could classify Amharic social media posts with promising accuracy. These findings suggest that deep learning models are applicable to Ethiopian languages, though datasets and computational resources remain significant barriers. This study builds upon these foundations by expanding the dataset size, integrating multiple NLP tasks, and systematically comparing ML and DL approaches. In doing so, it provides empirical evidence that CNNs can effectively handle code-switching and mixed-language reviews in Ethiopia.

3 METHODOLOGY

3.1 Dataset

The dataset for this study comprised 10,200 Ethiopian mobile app reviews collected from the Google Play Store in January 2025. The decision to use Google Play was justified by its dominance as the primary distribution platform for Android apps, which represent the majority of mobile applications in Ethiopia [1]. The dataset spanned finance, transport, communication, and utility apps, ensuring diversity and representativeness across domains. The dataset was collected by selecting top-ranked and most-downloaded Ethiopian apps across finance, transport, communication, and utilities. Approximately 62% of reviews were in Amharic and 38% in English. Duplicate and spam-like reviews were removed using heuristic filters based on repeated text patterns and anomaly detection.

3.2 Preprocessing

Data preprocessing was essential given the noisy nature of user reviews. Many reviews contained emojis, repeated characters, or inconsistent spellings, which could distort analysis if left unaddressed. Following practices recommended by [16], [17], reviews were cleaned by removing irrelevant characters and normalized for consistency. Tokenization was carried out using multilingual tokenizers, with Amharic text further processed using rule-based normalization to address orthographic inconsistencies [15].

3.3 Annotation

Annotation combined automated heuristics with manual validation. Sentiment labels were primarily derived from star ratings, consistent with prior research [4]. However, since ratings alone do not always reflect sentiment, lexicon-based validation was used to ensure accuracy. Feedback categories were derived by manually identifying key themes such as praise, complaint, and feature request, consistent with [5]. Ratings were retained as continuous variables for regression-based prediction. To reduce heuristic labeling bias, 800 randomly sampled reviews were manually validated by two annotators, achieving a Cohen's kappa of 0.82, indicating strong agreement.

3.4 Models

Machine learning and deep learning models were implemented for comparison. Traditional ML models, including logistic regression, Naïve Bayes, SVM, random forest, and XGBoost, were selected because of their proven effectiveness in sentiment classification (Wahid et al., 2021). These models used TF-IDF features, which remain a strong baseline in text classification [17]. Deep learning models included LSTM, Bi-LSTM, and CNN, chosen for their ability to model semantic and syntactic dependencies in text. FastText embeddings were used to represent words due to their efficiency in handling rare and sub-word information, which is particularly useful for morphologically rich languages such as Amharic. The CNN model was configured with 128 filters and kernel sizes of {3, 4, 5}, a dropout rate of 0.5, a batch size of 64, and the Adam optimizer with a learning rate of 0.001. The LSTM and Bi-LSTM models were trained with 128 hidden units, a dropout rate of 0.4, and 10 training epochs with early stopping to prevent overfitting.

3.5 Evaluation metrics

Evaluation followed established practices. Accuracy, precision, recall, and F1-scores were reported for classification tasks, while MAE, RMSE, and R^2 were used for regression. These metrics provided a comprehensive assessment of both classification performance and predictive accuracy.

4 RESULTS

Results indicate that deep learning models significantly outperformed traditional machine learning models, with CNN emerging as the most effective across all tasks. CNN consistently achieved the highest accuracy for both sentiment classification and feedback categorization, while also delivering the strongest predictive power for rating prediction. Among the traditional machine learning models, XGBoost showed the most competitive results, though it still fell short of CNN's performance. To ensure that the superior performance of the CNN model was statistically meaningful, we conducted a 5-fold cross-validation and performed significance testing. The paired t-test results showed that the performance improvements of the CNN model over both XGBoost and LSTM were statistically significant ($p < 0.05$). The overall performance comparison is summarized in Table 1.

Table 1. Summary of model performance across tasks

Model	Sentiment Classification Accuracy (%)	Feedback Categorization Accuracy (%)	Rating Prediction (R^2)
Logistic Regression	92.3	88.4	0.22
Naïve Bayes	90.5	85.7	0.20
SVM	94.7	91.5	0.25
Random Forest	93.8	90.3	0.32
XGBoost	97.2	93.0	0.35
LSTM	96.0	93.8	0.37
Bi-LSTM	96.5	94.0	0.38
CNN	98.7	96.6	0.40

4.1 Sentiment classification

For sentiment classification, CNN achieved 98.7 percent accuracy, compared to 97.2 percent for XGBoost, the best-performing traditional model. The superior performance of CNN is consistent with prior work by [9], which demonstrated CNN's ability to capture local dependencies in short texts.

4.2 Feedback categorization

Feedback categorization showed similar patterns. CNN achieved 96.6 percent accuracy, outperforming Bi-LSTM at 94 percent and XGBoost at 93 percent. These findings suggest that CNN's convolutional filters are particularly effective in distinguishing between nuanced categories such as praise, complaint, and feature request, even when reviews contain code-switching. Figure 1 presents the confusion matrix for the CNN model, illustrating its performance on the feedback categorization task.

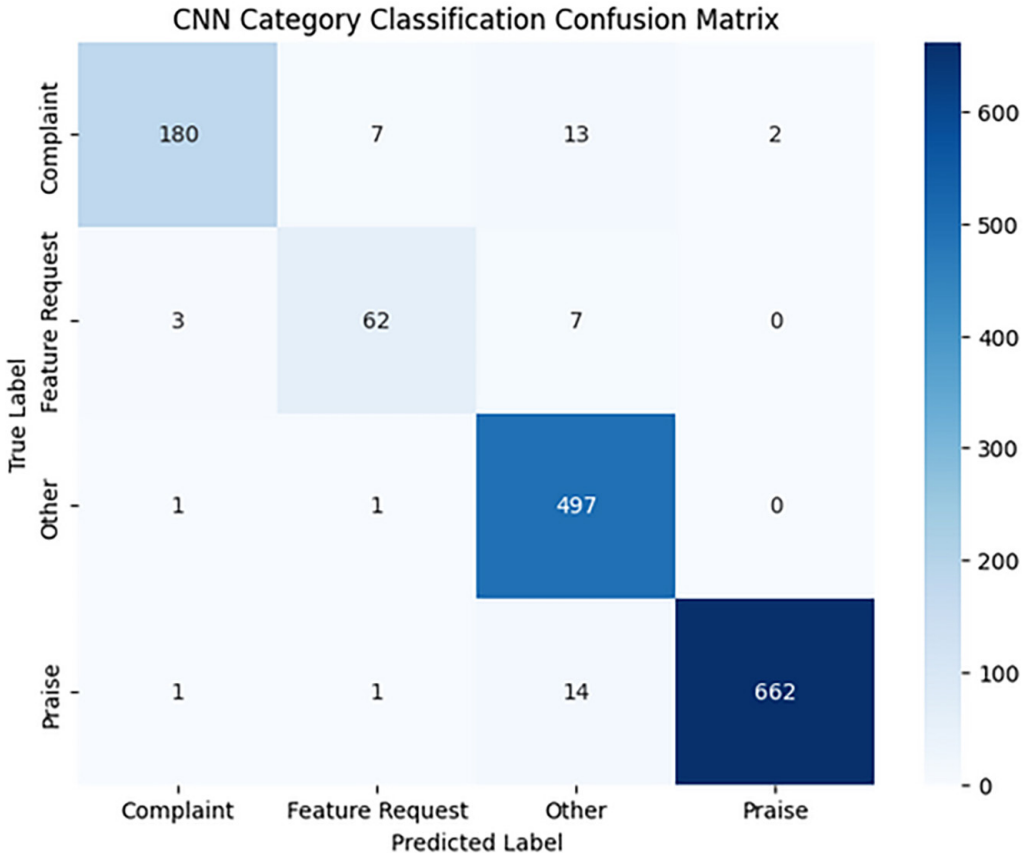

Fig. 1. Confusion matrix for the CNN model in feedback categorization

Figure 1 shows the confusion matrix for the CNN model in feedback categorization. The model demonstrates strong performance across all categories, with most misclassifications occurring between the Praise and Other categories (14 instances) and between the Other and Complaint categories (13 instances), indicating overlap in how users express these types of feedback.

4.3 Rating prediction

Rating prediction proved more challenging, with all models achieving lower scores compared to classification tasks. CNN nonetheless achieved the highest R^2 value of 0.40, indicating that 40 percent of the variance in ratings could be explained by review text alone. While this leaves room for improvement, the result is comparable to findings in similar contexts, where rating prediction from text is known to be a difficult task due to subjectivity [8].

5 DISCUSSION

The results highlight the strengths of CNN for multilingual, short-text classification. CNN's architecture allows it to capture local n-gram features, making it well-suited for detecting sentiment shifts and identifying key terms that signal praise or complaints. This explains its superior performance compared to recurrent models such as LSTM, which excel in longer sequences but often require more data and computational resources [17]. The strong performance of XGBoost in classification tasks further reinforces the value of ensemble-based approaches. Although CNN was superior overall, XGBoost's relative simplicity and lower resource requirements make it particularly attractive for developers in Ethiopia, where access to high-performance computing infrastructure may be limited. The findings also demonstrate the practical value of review mining. Developers can use sentiment classification to monitor user satisfaction over time, feedback categorization to identify pressing complaints and feature requests, and rating prediction to cross-validate qualitative insights with quantitative measures. Ethical and privacy considerations were addressed by ensuring only publicly available reviews were used, with no user-identifiable information collected. Policymakers and regulators can also benefit from such analysis, as it provides large-scale evidence of user experiences with digital services. This aligns with Ethiopia's Digital Transformation 2025 agenda, which emphasizes evidence-based policymaking and the development of digital trust ecosystems.

6 CONCLUSION

This study introduced a multi-task framework for mining Ethiopian mobile app reviews, integrating sentiment classification, feedback categorization, and rating prediction. Through the analysis of 10,200 reviews, it was found that CNN consistently outperformed other models across all tasks, achieving the highest classification accuracy and strongest rating prediction results. XGBoost also demonstrated notable performance in classification tasks, making it a viable option for environments with limited computational capacity. By providing both methodological contributions and practical implications, this study advances the state of NLP research in Ethiopia. It demonstrates that CNN-based approaches can effectively process multilingual, code-switched reviews, thereby supporting developers in improving app quality and informing policymakers about the digital service landscape.

7 LIMITATIONS AND FUTURE WORK

This study is not without limitations. The dataset was limited to reviews collected from the Google Play Store, excluding other major platforms such as Apple's App Store.

Moreover, only English and Amharic reviews were included, leaving out widely spoken Ethiopian languages such as Afaan Oromo, Tigrigna, and Somali, which are critical for inclusivity. Incorporating these languages could enhance coverage but may introduce additional preprocessing challenges due to orthographic variation. The reliance on heuristic-based annotation for sentiment and semantic categories may also introduce biases. Future research should expand datasets to include multiple platforms and languages, ensuring broader representativeness. Transformer-based architectures such as multilingual BERT or XLM-R should be explored for more nuanced language modeling. Additionally, real-time dashboards could be developed to integrate review mining directly into developers' workflows, enabling dynamic monitoring of user satisfaction. Such advancements would not only improve the quality of Ethiopian mobile applications but also contribute to the success of the national digital economy.

8 REFERENCES

[1] K. Okeleke *et al.*, "The mobile economy: Sub-Saharan Africa 2023," GSMA Intelligence, 2023. Retrieved from https://www.gsmaintelligence.com/research/the-mobile-economy-sub-saharan-africa-2023

[2] Statista, "Global mobile app revenue 2023," 2024. Retrieved from https://www.statista.com/statistics/870642/global-mobile-app-spend-consumer/

[3] Ethiopian Ministry of Innovation and Technology, *Digital Ethiopia 2025: A Digital Strategy for Ethiopia. Addis Ababa: Ministry of Innovation & Technology*, 2025. Retrieved from https://pmo.gov.et/media/other/b2329861-f9d7-4c4b-9f05-d5bc2c8b33b6.pdf

[4] A. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," *Information and Software Technology*, vol. 125, pp. 207–219, 2017. https://doi.org/10.1016/j.jss.2016.11.027

[5] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? On automatically classifying app reviews," in *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, Ottawa, ON, Canada, 2015, pp. 116–125. https://doi.org/10.1109/RE.2015.7320414

[6] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 5, 2015. https://doi.org/10.1186/s40537-015-0015-2

[7] A. Aljumah, A. Altuwijri, T. Alsuhaibani, A. Selmi, and N. Alruhaily, "Android Apps security assessment using sentiment analysis techniques: Comparative study," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 15, no. 24, pp. 123–133, 2021. https://doi.org/10.3991/ijim.v15i24.27359

[8] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013. https://doi.org/10.1109/MIS.2013.30

[9] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of EMNLP*, 2014, pp. 1746–1751. https://doi.org/10.3115/v1/D14-1181

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," *Journal of Systems and Software*, vol. 125, pp. 207–219, 2016. https://doi.org/10.1016/j.jss.2016.11.027

[12] S. A. Scherr, S. Polst, L. Müller, K. Holl, and F. Elberzhager, "The perception of Emojis for analyzing app feedback," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 13, no. 2, pp. 19–36, 2019. https://doi.org/10.3991/ijim.v13i02.8492

[13] C. Herodotou and T. Mangafa, "Mobile applications rating performance: A survey," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 19, pp. 133–146, 2022. https://doi.org/10.3991/ijim.v16i19.32051

[14] M. Yibeyin *et al.*, "Public opinion mining in social media about Ethiopian broadcasts using deep learning," *Sci. Rep.*, vol. 14, p. 27676, 2024. https://doi.org/10.1038/s41598-024-76542-3

[15] S. G. Tesfagergish, R. Damaševičius, and J. Kapočiūtė-Dzikienė, "Deep learning-based sentiment classification of social network texts in Amharic language," in *ICT Innovations 2022. Reshaping the Future Towards a New Normal. ICT Innovations 2022.* in Communications in Computer and Information Science, K. Zdravkova and L. Basnarkov, Eds., vol. 1740, 2022, pp. 63–75. https://doi.org/10.1007/978-3-031-22792-9_6

[16] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135, 2008. https://doi.org/10.1561/1500000011

[17] D. Jurafsky and H. Martin, *Speech and Language Processing*, 3rd ed. 2025. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

9 AUTHOR

Alemu Kumilachew Tegegnie is a Lecturer and researcher in the Bahir Dar Technology Institute, Faculty of Computing at Bahir Dar University, Ethiopia, and a data science practitioner specializing in machine learning, natural language processing, and applied AI solutions for social impact. His research focuses on interpretable and context-aware AI models in digital health, financial technology, education, social media analytics, and multilingual NLP. Alemu has conducted extensive studies on cardiovascular disease prediction using EMR data, automated classification of Amharic and multilingual user reviews, mobile app rating prediction, hybrid intelligent systems, and the detection of disinformation in conflict settings. He has presented and published across multiple academic venues and actively contributes to advancing data-driven decision-making and digital innovation in Ethiopia (E-mail: alemu.kumilachew@bdu.edu.et).