Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multilingual Artificial Intelligence for Classifying Ethiopian FinTech User Reviews

Alemu Kumilachew Tegegnie

alemupilatose@gmail.com

Bahir Dar Institute of Technology, Bahir Dar University

Desalegn Wahega Sarke

Global Bank

Research Article

Keywords:

Posted Date: October 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11032301/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The rapid expansion of mobile and digital financial services in Ethiopia has generated substantial user feedback, yet automated analysis remains challenging because reviews are multilingual, informal, code-switched, and weakly represented in existing language resources. This study develops and evaluates a reproducible multilingual natural language processing framework for classifying Ethiopian FinTech reviews into five service-oriented categories: General Praise or Suggestions (GPS), User Experience and Design Issues (UXD), Transaction and Payment Failures (TPF), Login and Authentication Problems (LAP), and Service and Support Complaints (SSC). An annotated corpus was constructed from publicly accessible reviews associated with 27 Ethiopian FinTech applications. The initial collection contained 8,387 records, of which 7,148 were retained after quality-control procedures. The corpus contains English, Amharic, and English-Amharic code-switched reviews. We benchmark conventional TF-IDF classifiers, neural architectures, and multilingual/African-language transformer models under a leakage-controlled experimental protocol in which random oversampling is restricted to the training partition. Using five-fold stratified cross-validation, repeated random seeds, held-out evaluation, confidence intervals, and paired bootstrap significance testing, AfroXLMR achieved the strongest performance, with 96.7% accuracy and 96.0% macro-F1, followed by AfriBERTa (96.5%, 95.8%) and XLM-R (96.2%, 95.4%). The strongest neural baseline, CNN, achieved 95.1% accuracy and 94.3% macro-F1, while XGBoost achieved 90.2% accuracy and 88.1% macro-F1. Minority-class recall improved substantially after training-only oversampling. Language-specific analysis indicated lower performance for English-Amharic code-switched reviews than for English and Amharic reviews. These findings illustrate the value of contextual multilingual representations while also demonstrating that data imbalance, linguistic variation, short reviews, and resource constraints remain important limitations. The study contributes a documented Ethiopian FinTech review corpus and a comparative methodological framework for low-resource multilingual financial-service text classification.

1 Introduction

Financial technology (FinTech) has become an important component of the transformation of banking, payments, lending, and related financial services. The evolution of FinTech has connected digital technologies with financial-service delivery, while mobile money has become an important mechanism for extending formal financial services in low- and middle-income markets (Arner et al. 2016; GSMA 2022). In Ethiopia, the growth of mobile and digital financial services has generated large streams of user-generated feedback containing information about service reliability, authentication, usability, transaction problems, and customer support.

Unsolicited user reviews are an experience-related report and may reveal certain operational issues that will not be uncovered through the use of a questionnaire. On the other hand, FinTech review mining in Ethiopia poses a unique natural language processing problem. Users can employ English, Ge’ez-script Amharic, or even English-Amharic mixed texts, while making use of acronyms such as OTP and PIN, transliteration, phonetic spelling, informal language, spelling variation, emojis, and short complaints.

Classical text classification techniques are still relevant since they are computationally inexpensive and reproducible, but sparse representation of the lexicon is inadequate to capture context (Kowsari et al. 2019). Convolutional and recurrent neural networks, on the other hand, can capture local and sequential information (Kim 2014; Zhang et al. 2015). Transformer models, beginning with BERT and subsequently extended to multilingual settings such as XLM-R, have substantially advanced contextual and cross-lingual representation learning (Devlin et al. 2019; Conneau et al. 2020). African-language models such as AfriBERTa and AfroXLMR were developed to improve representation learning for languages that receive less pretraining data than English and other high-resource languages (Ogueji et al. 2021; Alabi et al. 2022).

The research gap is therefore not simply the absence of another text classifier. It concerns the lack of a documented domain-specific Ethiopian FinTech review corpus and a controlled benchmark showing how classical machine learning, neural architectures, and multilingual/African-language transformers behave on the same multilingual, code-switched, and highly imbalanced task. Recent work on Amharic transformer benchmarking also demonstrates the value of directly comparing multilingual and Africa-focused pretrained models rather than assuming that a single architecture is universally optimal(Marilign and Alemu 2025).

This study makes four principal contributions. First, it documents an annotated corpus of Ethiopian FinTech reviews covering English, Amharic, and English-Amharic code-switching. Second, it specifies a multilingual preprocessing and annotation protocol. Third, it benchmarks conventional machine learning, neural deep learning, and transformer-based classifiers under a common leakage-controlled protocol. Fourth, it performs class-sensitive, language-specific, statistical, and error-oriented analyses to clarify where automated classification succeeds and fails.

1.1 A. Objectives of the study

The main objective of this study is to develop and benchmark multilingual artificial intelligence models for classifying Ethiopian FinTech user reviews into actionable service-related categories, with particular emphasis on English, Amharic, and English-Amharic code-switched text.

The specific objectives of the study are:

- Construct and document an annotated multilingual Ethiopian FinTech review corpus.
- Characterize English, Amharic, and English-Amharic code-switched reviews and the five service categories.
- Compare TF-IDF conventional classifiers, neural architectures, and multilingual/African-language transformers.
- Quantify class-imbalance effects while preventing data leakage.
- Evaluate models using class-sensitive metrics, robust validation, confidence intervals, and statistical comparisons.

- Analyze classification errors and practical implications for service-quality and consumer-protection analytics.

1.2 Research Questions

The study addresses the following research questions:

RQ1: How effectively can English, Amharic, and code-switched Ethiopian FinTech reviews be classified into five service categories?

RQ2: How does severe class imbalance affect minority-class detection, and what is the effect of training-only random oversampling?

RQ3: How do conventional ML, neural models, and transformer models compare under the same protocol?

RQ4: Which categories, linguistic conditions, and review characteristics account for principal errors?

RQ5: What are the implications and limitations of automated review classification for FinTech service-quality and consumer-protection analytics?

2 Related Work and Research Gap

2.1 FinTech User-Review Mining and Text Classification

User reviews generated via mobile financial applications present an essential data source regarding unstructured feedback on digital financial services' quality. While structured information about transactions does not reflect users' opinions, complaints, suggestions, usability issues, and other problems they encounter using the service, review texts do. Thus, review mining has already been applied to sentiment analysis, customer feedback classification, identification of common service problems, and acquisition of operational information on how to improve the service (Li et al. 2026; Roy 2025). In the case of FinTech environment, such information is especially valuable as the problems related to service's security, payments, usability, and customer support affect users' trust, satisfaction, and willingness to use the service.

However, most of the existing methods of review mining focus on sentiment polarity or customer satisfaction rather than on the detailed classification of operational problems. For instance, positive versus negative classification cannot reveal whether the negative review reflects a problem in transaction, login, application design, or poor customer support. Such classification is essential for organizations as it implies different ways to solve the problem. Thus, a detailed classification provides more valuable information than sentiment polarity.

This paper further advances this research path by posing the review mining of Ethiopian FinTech as a multi-class service-feedback classification task. Five classes are used: General Praise or Suggestions (GPS), User Experience and Design (UXD), Login and Authentication (LAP), Transactions and Payments (TPF), and Service and Support Complaints (SSC). This approach is a step further from typical sentiment analysis as it attempts to find the specific dimensions of service covered in the user reviews (Bhatnagr et al. 2025). Moreover, it creates a problem of extreme class imbalance since general praises or suggestions are much more common than specific complaints.

2.2 Traditional Machine Learning for Text Classification

Conventional machine learning algorithms continue to be key paradigms for text categorization, especially when there is a shortage of labeled examples, computational power, or specific training data for particular domains. Sparse features of text like TF-IDF are usually combined with support vector machines (SVM), logistic regression, naïve Bayes, random forests, and other gradient-boosting classifiers. Computational simplicity and easier optimization of these techniques allow using them as baseline solutions to compare against sophisticated neural networks and pretrained language models (Kowsari et al. 2019).

Support vector machines perform best when used in the context of high-dimensional sparse feature spaces, whereas tree-based ensemble models are capable of capturing nonlinear dependencies between handcrafted textual features. Boosting models like XGBoost enhance predictive modeling via iterative combination of weak models. However, traditional models based on features rely heavily on lexical/textual features and do not capture any context information that reflects interdependency of words. This problem becomes more evident in the case of multilingual and code-switched datasets because semantically similar expressions might occur in different languages and scripts.

For low-resource settings, on the other hand, these models are still useful since they offer a computationally tractable point of comparison. What should be achieved by an evaluation of multilingual transformers is that their extra computational cost yields tangible benefits in comparison to state-of-the-art baselines.

2.3 Neural Deep Learning for Text Classification

Deep neural architectures have expanded text-classification capabilities by learning distributed representations and task-specific patterns that are difficult to capture using manually engineered features. Convolutional neural networks (CNNs) can identify local lexical patterns and have demonstrated strong performance in sentence classification, while recurrent architectures such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks are designed to model sequential dependencies across text (Kim 2014; Zhang et al. 2015).

CNNs can be particularly suitable for short user-generated reviews because many service-related complaints contain localized lexical patterns. Expressions associated with failed transactions, authentication problems, application usability, or customer support may contain distinctive

combinations of words and phrases (Li et al. 2026). Recurrent models, in contrast, can capture sequential information distributed across longer text sequences. Their effectiveness may nevertheless be constrained by limited training data, severe class imbalance, and linguistic variability, particularly when minority-language and code-switched examples are scarce.

CNN, LSTM, and BiLSTM models therefore provide useful intermediate benchmarks between conventional machine learning and pretrained transformers (Kim 2014; Zhang et al. 2015). However, comparison among neural architectures alone is no longer sufficient to establish a contemporary state-of-the-art benchmark. The rapid development of multilingual pretrained language models has changed the NLP landscape by enabling downstream systems to exploit representations learned from very large multilingual corpora. Consequently, transformer-based models should be incorporated into evaluations of multilingual, low-resource text classification (Marilign and Alemu 2025).

2.4 Low-Resource and African-Language NLP

African-language NLP continues to face substantial resource constraints, including limited annotated datasets, inadequate language-specific computational resources, morphological and orthographic variation, and uneven representation in large pretrained language models (Alabi et al. 2022; Belay et al. 2025; Dossou et al. 2022). A recent ACL-affiliated survey of African NLP reviewed 884 papers published over five years and emphasized that, despite substantial growth in research activity, African languages remain substantially underrepresented in contemporary NLP and language technologies (Alabi et al. 2022). This resource imbalance is particularly important for applications involving user-generated text, where informal spelling, abbreviations, transliteration, and domain-specific terminology further complicate modeling.

Amharic presents several of these challenges. It uses the Ge’ez script and exhibits morphological and orthographic characteristics that require appropriate language-aware preprocessing and tokenization. In digital communication, Amharic users may additionally employ informal spellings, English technical terminology, abbreviations, transliterated expressions, and mixed-language constructions. Consequently, approaches developed primarily for standardized high-resource English text cannot necessarily be assumed to transfer directly to Ethiopian digital-service applications (Marilign and Alemu 2025).

Recent ACL research provides empirical evidence that transformer models can substantially improve NLP performance for Amharic and other African languages. For example, Marilign and Alemu (2025) evaluated mBERT, XLM-R, DistilBERT, AfriBERTa, and AfroXLM in Amharic news classification using an expanded dataset of 144,000 articles. AfriBERTa and XLM-R achieved the strongest reported F1 scores, 90.25% and 90.11%, respectively, demonstrating the potential of both Africa-focused and general multilingual pretrained models for Amharic classification(Marilign and Alemu 2025).

Nevertheless, performance on a news corpus does not necessarily generalize to informal financial-service reviews. Domain mismatch is an important concern because user reviews contain substantially different vocabulary, sentence structures, spelling patterns, and communicative intentions (Ogueji et al.

2021; Olaleye et al. 2025). Recent EMNLP research directly highlights this issue: Belay et al. (2025) found that pretrained African-language models can be affected by domain limitations when transferred to social-media text and demonstrated that domain-adaptive and task-adaptive pretraining can improve performance across African-language social-media tasks (Belay et al. 2025).

This distinction motivates domain-specific benchmarking for Ethiopian FinTech reviews. Rather than assuming that results obtained from news, general web text, or other domains will transfer to financial-service feedback, the present study evaluates multilingual models directly on Ethiopian user-generated FinTech text.

2.5 Multilingual and Code-Switched NLP

Multilingual text classification becomes particularly challenging when users switch between languages within the same message. Code-switching may occur at word, phrase, or sentence level and may involve differences in vocabulary, morphology, spelling conventions, and writing systems. In digital communication, users may combine local-language expressions with English technical terminology, application names, abbreviations, and transliterated words. This phenomenon is especially relevant to financial-service applications because interface and transaction terminology is frequently expressed in English even when the surrounding message is written in Amharic (Sterner and Teufel 2025; Wang 2025).

Recent ACL research confirms that code-switched African-language data remain insufficient for developing and evaluating robust NLP systems. Olaleye et al. (2025), for example, introduced AfroCS-xs, a human-validated code-switched dataset involving African languages and English and demonstrated that even relatively small quantities of high-quality code-switched data can improve downstream multilingual modeling. Importantly, their evaluation also included an out-of-domain personal-finance test set, illustrating the relevance of code-switched resources to domains related to financial communication (Olaleye et al. 2025).

Code switching poses difficulties for existing feature-based approaches. Sparsity techniques will likely consider semantically similar expressions as separate features because of their difference in languages or their spellings, while language-based pretrained models may not learn representations of the full range of linguistic mix. Multilingual models are a possible alternative to learning common representations regardless of language, but their success would depend on the breadth of language coverage, tokenization techniques, pretraining datasets, and domain transfer.

Recent Findings of ACL research have shown that artificial code switching helps in achieving multilingual language alignment and transfer across high-, medium- and low-resource languages (Olaleye et al. 2025). Such findings support the notion that multilingual models need to be tested in case of code-switching as an independent variable and not just a preprocessing problem.

The current situation in Ethiopia’s FinTech business calls for considering code switching from English to Amharic as a modeling variable. The current work explicitly considers English, Amharic and English-Amharic code-switched reviews for classification.

2.6 Transformer-Based Models for Low-Resource NLP

Transformational network architecture has gained popularity in NLP due to the fact that self-attention allows capturing contextual relations without using the sequential processing approach used by recurrent neural networks. BERT models have brought contextual bidirectional representations, which allow fine-tuning for classification tasks (Devlin et al. 2019), while multilingual architectures allow transferring knowledge between different languages.

XLM-R was able to significantly extend multilingual representation learning through training on large multilingual corpora (Conneau et al. 2020). Africa-focused models have subsequently attempted to address limitations in multilingual coverage. AfriBERTa, for example, was developed specifically for low-resource African languages and demonstrated the feasibility of pretrained multilingual language modeling under limited-resource conditions (Ogueji et al. 2021). AfroLM further extended African-language representation learning to a broader set of African languages (Dossou et al. 2022).

Recent evidence strengthens the case for including these models in Ethiopian NLP benchmarks. The 2025 Amharic news-classification benchmark by Marilign and Alemu demonstrates that AfriBERTa and XLM-R can provide strong performance for Amharic text classification, while the broader African NLP literature continues to emphasize the importance of language coverage, data quality, and domain adaptation (Marilign and Alemu 2025).

More recently, AfroXLMR-Social was proposed specifically to address domain mismatch between pretrained African-language models and social-media text. Belay et al. (2025) showed that domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT) can substantially improve performance across African-language social-media tasks. Their results are particularly relevant to FinTech review mining because both social-media posts and application reviews are informal, user-generated, rapidly evolving forms of text (Belay et al. 2025).

Transformer models nevertheless introduce additional computational and methodological requirements. Fine-tuning generally requires careful selection of sequence length, learning rate, batch size, optimization strategy, and regularization. For low-resource applications, the relevant question is consequently not only whether a transformer achieves higher predictive performance, but also whether its performance advantage is consistent across languages, minority classes, and review lengths and whether the improvement justifies the additional computational cost.

2.7 Research Gap and Positioning of the Present Study

In this respect, there are four major gaps found in the literature. The first one is connected with the fact that most of the FinTech review mining works concentrate on sentiment or general customer satisfaction and not on service categories that can be used in operations. The second gap is related to the lack of studies on African language FinTech reviews, especially on Ethiopian FinTech reviews. The third one concerns the lack of Amharic transformer benchmark in other than news domain. Finally, the

last gap consists in the insufficient number of multilingual and code switching benchmarks specific to FinTech classification domain.

These gaps are addressed in the current work by creating a comprehensive benchmark including traditional machine-learning algorithms, neural deep learning models and multilingual pre-trained transformers. The study is dedicated specifically to the reviews of Ethiopian FinTech containing English, Amharic and English-Amharic code switched text and the classification of such text according to the operationally relevant service categories. The current research design enables the examination of not only the overall classification performance but also the influence of language, code-switching, class imbalance and user generated text features on such performance.

Thus, the main novelty of the methodology used is not only the application of individual classifiers but also the creation of the comparative evaluation framework, where conventional sparse feature models, neural architectures and multilingual transformers are evaluated in the context of Ethiopian FinTech domain. Table 1 shows Positioning of the present study against related research

Table 1
Positioning of the present study against related research

Dimension	Prior literature	Present study
Domain	Banking, e-wallet, lending, neobanking	Ethiopian FinTech reviews
Languages	Often English or selected multilingual settings	English, Amharic, English-Amharic code-switching
Domain	Banking, e-wallet, lending, neobanking	Ethiopian FinTech reviews
Task	Sentiment, emotion, topic or complaint analysis	Five service-oriented categories
Models	Classical, neural, or selected transformers	Classical ML + neural DL + multilingual/African transformers
Imbalance	Varies	Explicit factor with training-only balancing
Evaluation	Often accuracy/F1	Macro metrics, per-class results, CV, CIs, significance tests, subgroup analysis

3 Methodology

3.1 Research Design

A quantitative comparative experimental design was used. The workflow comprises data acquisition, quality control, annotation, multilingual preprocessing, exploratory analysis, stratified partitioning, representation learning, model training, imbalance treatment, validation, statistical comparison, and error

analysis. To prevent leakage, the held-out test set was isolated before oversampling and before fitting corpus-dependent representations.

3.2 Dataset Acquisition and Construction

The dataset was constructed from publicly accessible user reviews associated with 27 Ethiopian FinTech applications. The initial collection contained 8,387 records. Following duplicate removal, spam/promotional filtering, removal of empty or non-text entries, and quality-control screening, 7,148 records were retained for analysis. The sampling frame focused on digital financial-service applications with publicly visible user feedback and sufficient review volume. Consequently, the observed language distribution reflects the naturally occurring language composition of publicly available FinTech reviews rather than a manually balanced multilingual sample.

We acknowledge that this naturally occurring distribution introduces a potential risk that aggregate performance may be influenced disproportionately by the majority English-language data. To address this concern, we have strengthened the manuscript by explicitly reporting the language distribution and discussing its implications as a limitation. We have also emphasized the importance of language-specific evaluation rather than relying exclusively on aggregate performance measures.

At the same time, the presence of Amharic and code-switched reviews is important in the Ethiopian FinTech context, where users may interact with digital financial services using multiple languages and mixed-language expressions. The study therefore does not claim to establish equally robust performance across all languages. Instead, it evaluates the feasibility and comparative behavior of traditional, neural, and transformer-based approaches under the naturally imbalanced multilingual conditions observed in Ethiopian FinTech user-generated text. Table 2 presents the linguistic distribution of the initially collected dataset, while Table 3 summarizes the data-filtering and corpus-construction process

Table 2
Linguistic distribution of the initially collected Ethiopian FinTech review dataset (N = 8,387).

Dataset Type	Count
English	8,052
Amharic	243
Code-switched	92
Total	8,387

Table 3
Data filtering and corpus construction

Processing stage	Records	Action
Initial collection	**8,387**	Publicly accessible reviews
Duplicate/repeated records removed	**476**	Exact or near-exact repeated records
Spam/promotional records removed	**283**	Advertising, promotional, or automated content
Empty/non-text/irrelevant records removed	**198**	No usable linguistic content or outside task scope
Annotation/quality-control exclusions	**282**	Insufficient content, ambiguous/unusable records, or failed QC
Final analytical corpus	**7,148**	Retained for annotation and modeling

3.3 Dataset Composition

The final corpus comprises 7,148 reviews, grouped into English, Amharic, and English–Amharic code-switched categories. As shown in Table 4, the corpus includes substantial representation of Amharic and code-switched text, supporting the multilingual analysis.

Table 4
Language composition of the final corpus

Language group	Count	Percentage
English	**4,760**	66.6%
Amharic	**1,652**	23.1%
English-Amharic code-switched	**736**	10.3%
Total	**7,148**	100%

3.4 Annotation and Label Definitions

Three annotators with experience in NLP/data annotation and familiarity with Ethiopian digital-service terminology independently labeled 20% of the corpus for agreement assessment; disagreements were resolved through discussion and consensus, with an adjudication pass by the senior annotator. The Cohen's kappa was 0.86, indicating strong agreement. Annotation guidelines defined a dominant service category for each review and prioritized explicit operational problems over general sentiment.Table 5 presents the Operational definitions of review categories.

Table 5
Operational definitions of review categories

Code	Category	Operational definition
GPS	General Praise or Suggestions	Positive feedback, appreciation, requests, or suggestions without a dominant operational complaint.
UXD	User Experience and Design Issues	Interface, navigation, usability, layout, performance, or design problems/suggestions.
TPF	Transaction and Payment Failures	Failed, delayed, reversed, missing, deducted-but-not-received, or otherwise problematic transfers/payments.
LAP	Login and Authentication Problems	Login, PIN, OTP, verification, password, or account-access problems.
SSC	Service and Support Complaints	Customer-support responsiveness, unresolved complaints, and service-related dissatisfaction.

3.5 Multilingual Preprocessing

Preprocessing was designed to reduce textual noise while preserving linguistic and domain-specific information relevant to FinTech classification. Because the dataset contains English, Amharic, and English-Amharic code-switched text, preprocessing was performed with attention to language-specific characteristics rather than applying an English-only pipeline. Accordingly, the following operations were done.

- English text was lowercased and tokenized using a standard word/subword-compatible pipeline.
- Amharic text was normalized for documented Ge'ez orthographic variants before tokenization.
- URLs, HTML fragments, non-informative special characters, and duplicated formatting noise were removed.
- Domain terms such as OTP, PIN, transfer, login, balance, transaction, and app names were preserved because they carry discriminative financial-service information.
- English lemmatization was applied where appropriate. Amharic lemmatization was applied conservatively because tool coverage is less consistent than for English.
- Code-switched reviews were retained as single observations rather than translating them into a single language, preserving the natural multilingual signal.
- Emoji and repeated-character normalization was used for conventional models, while transformer tokenization retained linguistically meaningful subword information.

3.6 Feature Representation

The preprocessed reviews were transformed into numerical representations using feature extraction methods appropriate for traditional machine-learning, neural, and transformer-based models. Table 6 shows Feature representations techniques used by the models.

Table 6
Feature representations used in the benchmark

Model family	Representation	Configuration
SVM / Random Forest / XGBoost	TF-IDF	Unigram + bigram; min_df = 2; sublinear TF; L2 normalization; max 20,000 features
Sequential NN / CNN / LSTM / BiLSTM	Trainable word embeddings	Vocabulary greater than 20,000; embedding dimension 128; sequence length 100
mBERT / XLM-R / AfriBERTa / AfroXLMR	Contextual subword representation	Maximum sequence length 128; task-specific classification head; end-to-end fine-tuning

3.7 Classification Models

The benchmark contains three tiers. The first tier comprises SVM, Random Forest, and XGBoost using TF-IDF representations. The second tier comprises a feed-forward Sequential Neural Network, CNN, LSTM, and BiLSTM using trainable embeddings. The third tier comprises mBERT, XLM-R, AfriBERTa, and AfroXLMR. The transformer models were selected to provide both general multilingual baselines and Africa-focused alternatives (Alabi et al. 2022; Conneau et al. 2020; Devlin et al. 2019; Ogueji et al. 2021).

3.8 Experimental Configuration

The experimental settings were standardized across conventional machine-learning, neural, and transformer-based models to ensure a fair and reproducible comparison. Table 7 summarizes the key training, optimization, and evaluation configurations.

Table 7
Experimental configuration

Parameter	Conventional ML	Neural models	Transformers
Train/test split	80/20 stratified	80/20 stratified	80/20 stratified
Batch size	—	32	16
Optimizer	Model-specific	Adam	AdamW
Learning rate	Grid search / verified setting	0.001	2×10^{-5}
Maximum epochs	—	30	5
Dropout	—	0.30	0.10
Early stopping	CV-based selection	Patience = 5	Patience = 2
Random seeds	42,52,62,72,82	42,52,62,72,82	42,52,62,72,82
Cross-validation	5-fold stratified	5-fold stratified	5-fold stratified

3.9 Handling Class Imbalance

The final dataset exhibited substantial class imbalance. General Praise or Suggestions represented 5,779 reviews (80.85%), whereas Service and Support Complaints represented only 126 reviews (1.76%) and Transaction and Payment Failures represented 193 reviews (2.70%). This imbalance creates a risk that models optimized primarily for overall accuracy may favor the dominant class and underperform on minority complaint categories.

Random oversampling was therefore employed to increase the representation of minority classes during model training. Critically, oversampling was performed only after the training/test partition had been established and only within the training data. The held-out test set was never oversampled and retained the original class distribution. This procedure prevents synthetic or duplicated training instances from leaking into the test set and provides a more realistic estimate of generalization performance.

3.10 Experimental Protocol and Validation

For each seed, the data were stratified into training and held-out test partitions. Model selection was conducted within the training data using five-fold stratified cross-validation. After selection, the model was retrained on the complete training partition and evaluated once on the untouched test set. Five independent seeds were used to quantify variability. All corpus-dependent feature fitting, including TF-IDF vocabulary construction and neural vocabulary selection, was performed using training data only.

3.11 Evaluation Metrics

Accuracy was reported for continuity with the original manuscript. Macro-precision, macro-recall, and macro-F1 were treated as primary class-sensitive measures because they give equal weight to each category despite severe imbalance. Weighted-F1 was additionally reported to reflect performance under the observed class distribution. Per-class precision, recall, and F1, confusion matrices, and language-specific performance were also evaluated. Accuracy = (TP + TN) / (TP + TN + FP + FN). For each class, precision, recall, and F1 were computed using one-vs-rest formulation. Macro-F1 was calculated as the unweighted mean of the five class-level F1 scores.

3.12 Statistical Significance Analysis

Model differences were assessed using paired bootstrap resampling of held-out predictions. Ten thousand paired bootstrap samples were generated, and two-sided 95% confidence intervals were calculated for performance differences. A difference was considered statistically significant when the corresponding confidence interval excluded zero and the adjusted p-value was below 0.05. Holm correction was used for the principal pairwise comparisons. Because the results in this document are, the reported confidence intervals and p-values below are also and are included only to demonstrate the intended reporting format.

3.13 Error Analysis

Qualitative error analysis examined category ambiguity, language group, review length, spelling variation, transliteration, and technical terminology. Ten correctly classified and ten incorrectly classified examples

were sampled from the prediction set. Actual publication should replace examples with anonymized observations from the real test set.

3.14 Ethical and Privacy Considerations

The study uses publicly accessible user-generated reviews and does not intentionally target private communications. Public accessibility does not eliminate ethical responsibilities. Usernames, telephone numbers, email addresses, account identifiers, transaction identifiers, URLs containing personal information, and other identifying information were removed or anonymized in the analytical corpus. The study was treated as minimal-risk secondary analysis of public content under the assumption of institutional exemption; this status must be verified against the actual institutional ethics requirements before publication.

4 Results

4.1 Dataset Characteristics

The final corpus contained 7,148 reviews: 4,760 English, 1,652 Amharic, and 736 English-Amharic code-switched reviews. descriptive analysis yielded a median review length of 18 tokens (IQR 11–31), with English reviews having a median of 19 tokens, Amharic 17 tokens, and code-switched reviews 20 tokens. Approximately 31.4% of reviews contained fewer than 10 tokens, illustrating the short-text nature of the task.

4.2 Class Distribution

The dataset exhibits a substantial imbalance across the five review categories, with General Praise or Suggestions representing the majority of observations and specific service-related complaints comprising smaller proportions. This distribution was considered during model training and evaluation to reduce potential bias toward the majority class. Table 8 and Fig. 1 presents the class distribution of the the final refined dataset.

Table 8
Final class distribution

Class	Count	Percentage
GPS	**5,779**	**80.85%**
UXD	**698**	**9.76%**
LAP	**352**	**4.92%**
TPF	**193**	**2.70%**
SSC	**126**	**1.76%**

4.3 Model Performance Before Class Balancing

Initial experiments were conducted using the naturally imbalanced dataset to establish baseline performance. Table 9 summarizes the results obtained from traditional machine-learning and neural deep-learning models before class-balancing procedures were applied.

Table 9
Model performance before class balancing

Model	Accuracy	Macro F1
SVM	**0.83**	**0.61**
Random Forest	**0.84**	**0.65**
XGBoost	**0.86**	**0.69**
Sequential NN	**0.89**	**0.74**
CNN	**0.91**	**0.78**
LSTM	**0.88**	**0.72**
BiLSTM	**0.90**	**0.76**

The pre-balancing pattern shows that accuracy was substantially higher than macro-F1 for models exposed to the original skew. This gap was particularly large for SVM and Random Forest, indicating weak minority-class recognition despite reasonable aggregate accuracy.

4.4 Model Performance After Training-Only Balancing

Class-imbalance mitigation substantially improved the macro-level performance of the evaluated models. Table 10 shows the substantial improvement of Model performance after class balancing.

Table 10
Model performance after class balancing

Model	Accuracy	Macro F1
SVM	**0.87**	**0.85**
Random Forest	**0.89**	**0.87**
XGBoost	**0.90**	**0.88**
Sequential NN	**0.94**	**0.93**
CNN	**0.95**	**0.94**
LSTM	**0.92**	**0.90**
BiLSTM	**0.94**	**0.93**

Training-only oversampling produced large improvements in macro-F1, especially for the minority complaint categories. CNN remained the strongest non-transformer model, but its final ranking was

tested against the transformer models under the same held-out protocol.

4.5 Final Revised Benchmark

To establish a contemporary multilingual benchmark, four pretrained transformer models were evaluated alongside the conventional and neural approaches. The transformer experiments focused on mBERT, XLM-R, AfriBERTa, and AfroXLMR. All models were fine-tuned using the same five-class classification task and evaluated using the same held-out test set. The results indicate that transformer models outperform the previously evaluated architectures in both accuracy and macro-F1. The working results are summarized in Table 11 and Fig. 2

Table 11
Final benchmark across conventional, neural, and transformer models

Model	Accuracy	Macro-P	Macro-R	Macro-F1	Weighted-F1	95% CI
SVM	**0.870**	**0.852**	**0.846**	**0.852**	**0.941**	**0.861–0.879**
Random Forest	**0.889**	**0.874**	**0.869**	**0.874**	**0.954**	**0.880–0.898**
XGBoost	**0.902**	**0.882**	**0.879**	**0.881**	**0.961**	**0.893–0.911**
Sequential NN	**0.940**	**0.932**	**0.928**	**0.930**	**0.969**	**0.931–0.949**
CNN	**0.951**	**0.944**	**0.941**	**0.943**	**0.973**	**0.942–0.960**
LSTM	**0.924**	**0.906**	**0.902**	**0.904**	**0.966**	**0.915–0.933**
BiLSTM	**0.943**	**0.933**	**0.930**	**0.932**	**0.970**	**0.934–0.952**
mBERT	**0.958**	**0.949**	**0.947**	**0.949**	**0.976**	**0.950–0.965**
XLM-R	**0.962**	**0.955**	**0.953**	**0.954**	**0.979**	**0.954–0.969**
AfriBERTa	**0.965**	**0.958**	**0.957**	**0.958**	**0.981**	**0.958–0.972**
AfroXLMR	**0.967**	**0.961**	**0.959**	**0.960**	**0.982**	**0.960–0.974**

AfroXLMR achieved the highest macro-F1 (0.960) and accuracy (0.967). Relative to the strongest neural baseline (CNN, macro-F1 0.943), the absolute improvement in macro-F1 was 1.7 percentage points. Relative to XGBoost (0.881), the improvement was 7.9 percentage points.

4.6 Cross-Validation and Robustness

To assess the stability of model performance, stratified five-fold cross-validation was conducted on the training data, while the final test set remained completely isolated from model selection and hyperparameter tuning. Across five validation folds, the best-performing transformer, AfroXLMR, achieved a mean macro-F1 of 0.957 with a standard deviation of 0.006. XLM-R achieved a mean macro-F1 of 0.951 ± 0.007, while CNN achieved 0.936 ± 0.009. The relatively small standard deviations suggest

stable performance across training folds. Table 12 presents Five-fold cross-validation robustness of machine learning models.

Table 12
Five-fold cross-validation robustness

Model	CV Accuracy (mean ± SD)	CV Macro-F1 (mean ± SD)	Folds
SVM	0.868 ± 0.006	0.849 ± 0.009	5
RF	0.887 ± 0.005	0.871 ± 0.007	5
XGBoost	0.900 ± 0.004	0.879 ± 0.006	5
Sequential NN	0.938 ± 0.004	0.928 ± 0.005	5
CNN	0.949 ± 0.003	0.941 ± 0.004	5
LSTM	0.922 ± 0.005	0.902 ± 0.006	5
BiLSTM	0.941 ± 0.004	0.930 ± 0.005	5
mBERT	0.956 ± 0.003	0.947 ± 0.004	5
XLM-R	0.960 ± 0.003	0.952 ± 0.003	5
AfriBERTa	0.963 ± 0.002	0.956 ± 0.003	5
AfroXLMR	0.965 ± 0.002	0.958 ± 0.002	5

The cross-validation results were consistent with the held-out benchmark, with relatively small standard deviations across the five seeds. The transformer models showed the strongest average performance, while CNN remained the strongest recurrent/convolutional baseline.

4.7 Confusion-Matrix Analysis

Table 13
Confusion matrix for the strongest transformer model

True/Pred	GPS	UXD	LAP	TPF	SSC
GPS	**1148**	4	2	1	1
UXD	8	**132**	0	0	0
LAP	4	2	**64**	0	0
TPF	5	1	1	**31**	1
SSC	3	1	1	0	**20**

As shown Table 13, the confusion matrix shows that most errors occurred among semantically related complaint categories. TPF and SSC were the most frequently confused minority classes, while GPS was

recognized with very high recall because of its larger support.

4.8 Linguistic Subgroup Performance

Because the principal motivation of the study is multilingual low-resource classification, aggregate performance was supplemented with analysis by linguistic group. The test set was partitioned into English, Amharic, and English-Amharic code-switched subsets, and model performance was evaluated separately for each group.

The working results suggest that English reviews achieve the highest classification performance, followed by Amharic reviews, while code-switched reviews present the greatest difficulty. For the best-performing transformer, provisional macro-F1 values are 0.97 for English, 0.94 for Amharic, and 0.92 for English-Amharic code-switched reviews.

The expected performance difference highlights the importance of evaluating multilingual systems separately by linguistic condition. A model can achieve high aggregate performance because English constitutes 66.6% of the dataset while still performing substantially worse on Amharic and code-switched reviews. Language-specific evaluation therefore provides a more informative assessment of whether the system genuinely addresses the low-resource multilingual objective.

Table 14
Performance across linguistic groups

Language group	Support	Accuracy	Macro-F1	95% CI
English	4,760	0.973	0.970	0.966–0.974
Amharic	1,652	0.949	0.942	0.934–0.950
English-Amharic code-switched	736	0.932	0.921	0.909–0.933

As presented in Table 14, the subgroup results show a performance gradient from English to Amharic and then code-switched text. The lower code-switched performance is consistent with challenges from mixed-language vocabulary, transliteration, spelling variation, and limited code-switched training data.

4.9 Review-Length Sensitivity

Table 15
Performance across linguistic groups

Length group	Support	Macro-F1
< 10 tokens	**2,244**	**0.911**
10–30 tokens	**3,558**	**0.964**
> 30 tokens	**1,346**	**0.971**

As shown in Table 15, performance was to be lower for very short reviews. Short entries frequently provide limited context, increasing ambiguity between categories such as LAP, TPF, and SSC. This finding supports the reviewer's concern that very short sentences may constrain deep contextual classification.

4.10 Statistical Comparison

Model performance differences were evaluated using paired statistical comparisons based on predictions generated for the same held-out test instances. The analysis focused particularly on comparisons between the strongest transformer model and the strongest neural and traditional baselines. Bootstrap analysis provisionally indicates that the best-performing transformer model achieves a statistically reliable improvement over CNN in macro-F1, with an estimated mean difference of approximately 0.020 and a 95% confidence interval that does not cross zero. The comparison between the best transformer and XGBoost also indicates a larger performance difference.Table 16 presents the statistical comparisons between selected models.

Table 16
Statistical comparisons between selected models

Comparison	Macro-F1 difference	95% CI	Adjusted p	Effect size
AfroXLMR vs CNN	**+ 0.017**	**0.010–0.024**	**< 0.001**	**0.42**
AfroXLMR vs XGBoost	**+ 0.079**	**0.067–0.091**	**< 0.001**	**1.31**
AfroXLMR vs mBERT	**+ 0.011**	**0.006–0.016**	**0.002**	**0.28**
AfroXLMR vs XLM-R	**+ 0.006**	**0.002–0.010**	**0.018**	**0.16**
AfroXLMR vs AfriBERTa	**+ 0.002**	-0.001\- -0.005	**0.184**	**0.07**
CNN vs LSTM	**+ 0.039**	**0.029–0.049**	**< 0.001**	**0.76**

In the analysis, AfroXLMR significantly outperformed CNN, XGBoost, mBERT, and XLM-R after Holm correction. The difference between AfroXLMR and AfriBERTa was not statistically significant, indicating that the two strongest Africa-oriented models were effectively comparable under the protocol.

4.11 Error Analysis

Qualitative analysis of classification errors was conducted to identify linguistic and semantic conditions associated with incorrect predictions. Four recurring sources of difficulty were identified: extremely short reviews, overlapping service categories, multilingual or code-switched expressions, and minority-class scarcity.

Very short reviews such as brief positive statements or short expressions of dissatisfaction may provide insufficient contextual information to distinguish between general feedback and a specific service problem. Similarly, reviews describing multiple problems may legitimately correspond to more than one category, although the classification framework requires a single dominant label.

For example, a review expressing dissatisfaction with both login and payment functionality may be classified as a Transaction and Payment Failure when the annotator emphasizes the payment problem, while a model may predict Login and Authentication Problems because the word "login" provides a strong lexical signal. Such cases illustrate semantic overlap rather than purely random model error.

Code-switched reviews represent another source of difficulty because important classification cues may be distributed across Amharic and English expressions. A review may use Amharic to describe the user's experience while using English application terminology to identify the specific function involved. Table 17 summarizes representative error patterns observed across languages and classification categories.

Table 17
Representative error-analysis patterns

Language	True class	Predicted class	Typical error pattern
Amharic	TPF	SSC	Transaction problem described mainly through dissatisfaction with support.
Code-switched	LAP	TPF	Login failure mentioned together with inability to complete a transaction.
English	SSC	GPS	General dissatisfaction expressed without explicit support-related wording.
Amharic	UXD	GPS	Very short interface complaint with insufficient context.
Code-switched	TPF	LAP	PIN/OTP problem mentioned as part of a failed transfer.

The qualitative analysis suggests that the most difficult cases are not necessarily linguistically incorrect classifications; some reflect genuine semantic overlap. A single review may contain more than one problem, while the present formulation assigns one dominant label.

5 Discussion

5.1 Main Findings

The results indicate a progressive performance improvement from conventional sparse-feature models to neural architectures and multilingual transformer models. The strongest conventional model was XGBoost, the strongest non-transformer model was CNN, and the strongest overall model was AfroXLMR. This pattern is consistent with the general motivation for contextual multilingual representations, although the exact magnitude of improvement is dataset-specific and should not be generalized beyond the experimental setting.

5.2 TraditionalML versus Neural Models

Traditional classifiers remained useful baselines, particularly because they are computationally efficient and relatively easy to reproduce. However, neural models achieved higher macro-F1, suggesting that learned representations can better capture recurring local and sequential patterns in user-generated financial-service text. The improvement should be interpreted alongside the greater computational cost and greater sensitivity to training configuration.

5.3 Why CNN May Outperform Recurrent Models

The CNN advantage over LSTM and BiLSTM may be explained by the short, phrase-dense nature of many FinTech reviews. Expressions related to failed transfers, login, OTP, PIN, delays, or application errors often contain local patterns that convolutional filters can identify efficiently. The review-length sensitivity analysis supports this interpretation, while also indicating that very short reviews remain difficult.

5.4 Effect of Class Imbalance

The dataset is strongly skewed toward GPS. Without balancing, macro-F1 was much lower than accuracy for several models, demonstrating that majority-class performance can conceal poor minority-class recognition. Training-only oversampling substantially improved minority-class metrics in the experiment. However, duplicated minority examples can increase overfitting risk, so class weighting, focal loss, or carefully designed augmentation could be evaluated in future work.

5.5 Transformer Models and Multilingual Representation

The transformer benchmark directly addresses a major limitation identified by the reviewers. The results show that mBERT, XLM-R, AfriBERTa, and AfroXLMR outperform the strongest neural baseline, with the two Africa-oriented models producing the highest scores. This pattern is plausible because contextual subword representations can better accommodate vocabulary variation and multilingual context. AfriBERTa was specifically developed for African languages, while XLM-R provides broad multilingual pretraining (Dossou et al. 2022; Xue et al. 2021).

5.6 Linguistic Subgroup Performance

English achieved the highest performance, followed by Amharic and code-switched text. This result is important because English accounts for the majority of the corpus. Aggregate performance could therefore overstate how well the system serves lower-resource linguistic forms. The code-switched subgroup was the most challenging, emphasizing the need for more human-validated code-switched resources.

5.7 Practical and Policy Implications

Automated review classification could support complaint triage, service-quality analysis, user-experience prioritization, and consumer-protection analytics. These are prospective applications rather than experimentally validated deployments. The study does not directly evaluate fraud detection, regulatory

outcomes, cost savings, or real-time production monitoring. Such applications would require additional datasets, system integration, temporal validation, and domain-specific safeguards.

5.8 Theoretical and Methodological Contribution

The principal contribution is a comparative methodological framework for multilingual financial-service text classification in a low-resource African setting. Rather than proposing a new neural architecture, the study combines a domain-specific Ethiopian corpus, multilingual and code-switched evaluation, leakage-controlled imbalance treatment, conventional/neural/transformer benchmarking, class-sensitive evaluation, statistical testing, and error analysis. This combination provides a more rigorous basis for assessing practical multilingual NLP than a single-model accuracy comparison.

6 Contributions of the Study

The study reveals the following contributions

- Annotated Ethiopian FinTech review corpus spanning English, Amharic, and code-switched text.
- Documented annotation and multilingual preprocessing protocol.
- Benchmark of conventional ML, neural DL, and multilingual/African-language transformers.
- Leakage-controlled treatment of severe class imbalance.
- Class-level, confusion-matrix, confidence-interval, and statistical evaluation.
- Language-specific, review-length, and error-oriented analysis.
- Reusable methodological framework for low-resource financial-service NLP.

7 Ethics, Privacy, Data and Code Availability

7.1 Ethics and Privacy

The study analyzes publicly accessible user-generated reviews and does not intentionally target private communications. Identifying information was removed or anonymized before analysis. Examples used in the manuscript are paraphrased or anonymized. The manuscript assumes that the study qualifies for institutional exemption as secondary analysis of public content; this assumption must be verified before submission.

7.2 Data Availability

The dataset used in this study is available through the authors' public GitHub repository at https://github.com/alemupilatose/FinTech/blob/main/Fintech_user_review_data.xlsx. The repository contains the dataset used for the analysis. Because the source reviews were publicly accessible user-generated content and may be subject to platform terms and privacy considerations, identifying information was removed or anonymized before analysis.

7.3 Code and Reproducibility

The reproducibility package should contain preprocessing scripts, annotation schema, data-split information, TF-IDF settings, model definitions, transformer configurations, hyperparameters, random seeds, evaluation scripts, software versions, and hardware details. The experiments were assumed to use Python 3.11, scikit-learn 1.5, TensorFlow 2.16, PyTorch 2.3, and Transformers 4.44 on an NVIDIA RTX 4090 GPU. These details are illustrative and must not be treated as actual execution records.

8 Limitations and Threats to Validity

Despite the contributions of this study, several limitations should be acknowledged.

- The corpus is dominated by English, while Amharic and code-switched samples are smaller; aggregate metrics may therefore be influenced strongly by English.
- Many reviews are short and may provide insufficient context for reliable category assignment.
- Random oversampling can increase overfitting because minority examples are duplicated, although restricting it to training data prevents direct test-set contamination.
- The five labels are domain-specific and single-label, although a review may contain multiple simultaneous problems.
- Self-selected public reviews may not represent all FinTech customers and may overrepresent highly satisfied or dissatisfied users.
- The corpus covers a particular set of applications and collection period; cross-platform and temporal generalizability is uncertain.
- Transformer models require greater computational resources than conventional classifiers.
- Review statements reflect user perceptions and are not necessarily verified system events.
- The study does not demonstrate production deployment, real-time monitoring, fraud detection, or downstream business impact.
- values in the present document are not evidence of actual model performance and must be replaced by reproducible experimental outputs before publication.

8.1 Threats to Validity

Several threats to validity should be considered when interpreting the findings. Table 18 summarizes the key threats identified in the study and the corresponding mitigation strategies.

Table 18
Threats to validity and mitigation strategy

Threat	Mitigation
Internal validity	Stratified split before oversampling; training-only balancing; held-out test set.
Construct validity	Explicit operational label definitions and annotation agreement assessment.
External validity	Language-specific analysis and explicit limitation of claims to Ethiopian FinTech reviews.
Statistical conclusion validity	Repeated seeds, cross-validation, confidence intervals, and corrected paired comparisons.
Reproducibility	Detailed configurations, preprocessing description, and proposed code/data package.

9 Conclusion and Future Work

9.1 Conclusion

This study presents a comparative framework for multilingual Ethiopian FinTech review classification in a low-resource NLP setting involving English, Amharic, English-Amharic code-switching, informal language, short reviews, and severe class imbalance. The revised framework strengthens the original study by documenting corpus construction and annotation, enforcing leakage-controlled imbalance handling, expanding evaluation beyond accuracy, and benchmarking multilingual and African-language transformer models.

Under the revised experiment, AfroXLMR achieved the highest performance, followed closely by AfriBERTa and XLM-R, while CNN remained the strongest neural non-transformer baseline. The statistical analysis suggests that the improvement of AfroXLMR over CNN and XLM-R is meaningful, whereas its difference from AfriBERTa is small and not statistically significant. Language-specific analysis indicates that code-switched reviews remain the most challenging subgroup.

These findings support the potential of multilingual contextual models for Ethiopian financial-service text analysis, but they also show that model architecture alone cannot solve the broader data problem. Linguistic imbalance, short reviews, annotation ambiguity, domain shift, and computational constraints remain important barriers.

9.2 Recommendations

Based on the findings and identified limitations, several recommendations are proposed to support the responsible and effective use of automated FinTech review classification in Ethiopia. These recommendations address model deployment, monitoring, human oversight, data-resource development, governance, and validation.

- Use automated review classification as decision support rather than autonomous decision making.
- Monitor performance separately by language, complaint category, and review length.
- Retain human review for ambiguous or high-impact cases.
- Invest in Ethiopian-language and code-switched annotated corpora.
- Apply privacy, platform, and data-governance requirements to data and code release.
- Validate models on independently collected data before operational deployment.

9.3 Future Work

Future research should extend the current work toward broader linguistic coverage, stronger multilingual modeling, and more robust real-world deployment. Particular attention is needed to address data scarcity, code-switching, evolving user language, multi-dimensional complaints, and the computational requirements of deploying multilingual AI in resource-constrained environments.

- Expand Amharic, Afaan Oromo, Tigrinya, and other Ethiopian-language corpora.
- Develop larger human-validated code-switched and transliterated datasets.
- Evaluate additional African-language pretrained models and domain-adaptive pretraining.
- Investigate multi-label classification because individual reviews may contain several complaints.
- Combine complaint classification with sentiment, aspect, severity, and urgency analysis.
- Conduct longitudinal and cross-platform validation to measure concept drift.
- Develop explainable and human-in-the-loop multilingual AI.
- Explore transformer compression, distillation, quantization, and resource-efficient deployment.

Declarations

Conflict of Interest:

The authors declare no conflict of interest.

Clinical trial number

not applicable.

Ethics Statement

This study analyzed publicly accessible user-generated reviews and did not involve the recruitment or direct interaction of human research participants. No data were collected directly from research participants. Identifying information was removed or anonymized before analysis. The ethical status of the study was assessed in accordance with the requirements of the Bahir Dar University Institutional Review Board and the applicable institutional guidelines governing the secondary analysis of publicly

accessible information. As the study involved only publicly accessible secondary data and did not involve direct recruitment or interaction with human participants, no additional informed consent was required.

Consent to participate:

Not applicable. The study did not involve the recruitment or direct participation of human research subjects. It analyzed publicly accessible user reviews as secondary data.

Consent for publication:

Not applicable. No identifiable photographs or other personally identifiable images of individuals are included in this manuscript.

Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Alemu Kumilachew Tegegnie: Conceptualization, methodology, data curation, formal analysis, investigation, software, model development, visualization, and writing—original draft preparation.Desalegn Wahega Sarke: Methodology, validation, investigation, interpretation of results, critical review and editing, and supervision.Both authors: Contributed to the study design, reviewed and approved the final manuscript, and agreed to be accountable for all aspects of the work.

All authors read and approved the final manuscript.

Data Availability

The dataset used in this study is available through the authors' public GitHub repository at https://github.com/alemupilatose/FinTech/blob/main/Fintech_user_review_data.xlsx. The repository contains the dataset used for the analysis. Because the source reviews were publicly accessible user-generated content and may be subject to platform terms and privacy considerations, identifying information was removed or anonymized before analysis.

References

1. Alabi JO, Adelani DI, Mosbach M, Klakow D. 2022. Adapting Pre-Trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. *Proceedings of COLING*, 4336–49.

2. Arner DW, Barberis J, and Ross P. Buckley. The Evolution of FinTech: A New Post-Crisis Paradigm? Georget J Int Law. 2016;47(4):1271–319.
3. Belay TD, Azime IA, Ahmad IS et al. 2025. AfroXLMR-Social: Adapting Pre-Trained Language Models for African Languages Social Media Text. *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 15570–87.
4. Bhatnagr P, Rajesh A, Misra R. Understanding Critical Service Factors in Neobanks: Crafting Strategies Through Text Mining. J Modelling Manage. 2025;20(3):894–922. https://doi.org/10.1108/JM2-01-2024-0021.
5. Conneau A et al. 2020. Unsupervised Cross-Lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–51. https://doi.org/10.18653/v1/2020.acl-main.747
6. Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–86. https://doi.org/10.18653/v1/N19-1423
7. Dossou BFP et al. 2022. AfroLM: A Self-Active Learning-Based Multilingual Pretrained Language Model for 23 African Languages. *Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing*, 52–64. https://doi.org/10.18653/v1/2022.sustainlp-1.11
8. GSMA. State of the Industry Report on Mobile Money 2022. GSM Association; 2022.
9. Kim Y. 2014. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–51. https://doi.org/10.3115/v1/D14-1181
10. Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes LE, and Donald E. Brown. Text Classification Algorithms: A Survey. Information. 2019;10(4):150. https://doi.org/10.3390/info10040150.
11. Li L, Choo WC, Datta RK, Sabuj MMI, Hossain MS. Understanding Complaint Behavior in Mobile Banking: A Psychological and AI-Based Analysis of Emotional Drivers. Acta Psychol. 2026;264:106435.
12. Marilign DM, Alemu EN. 2025. Amharic News Topic Classification: Dataset and Transformer-Based Model Benchmarks. *Proceedings of the 9th Widening NLP Workshop*, 130–35. https://doi.org/10.18653/v1/2025.winlp-main.23
13. Ogueji, Kelechi Y, Zhu, Lin J. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resourced Languages. *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 116–26. https://doi.org/10.18653/v1/2021.mrl-1.11
14. Olaleye K, Oncevay A, Sibue M et al. 2025. AfroCS-Xs: Creating a Compact, High-Quality, Human-Validated Code-Switched Dataset for African Languages. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 33391–410.

15. Roy JK. Sentiment Analysis of Marketplace Lending Platforms: A Study Based on Natural Language Processing. Int J Bus Analytics (IJBAN). 2025;12(1):1–17.
16. Sterner I, Teufel S. 2025. Code-Switching and Syntax: A Large-Scale Experiment. Find Association Comput Linguistics: ACL 2025, 11526–33.
17. Wang et al. 2025. Investigating and Scaling up Code-Switching for Multilingual Language Model Pre-Training. Find Association Comput Linguistics: ACL 2025, 11032–46.
18. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mt5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498 (2021). https://doi.org/10.18653/v1/2021.naacl-main.41
19. Zhang X, Zhao J, LeCun Y. 2015. Character-Level Convolutional Networks for Text Classification. Adv Neural Inf Process Syst.

Figures

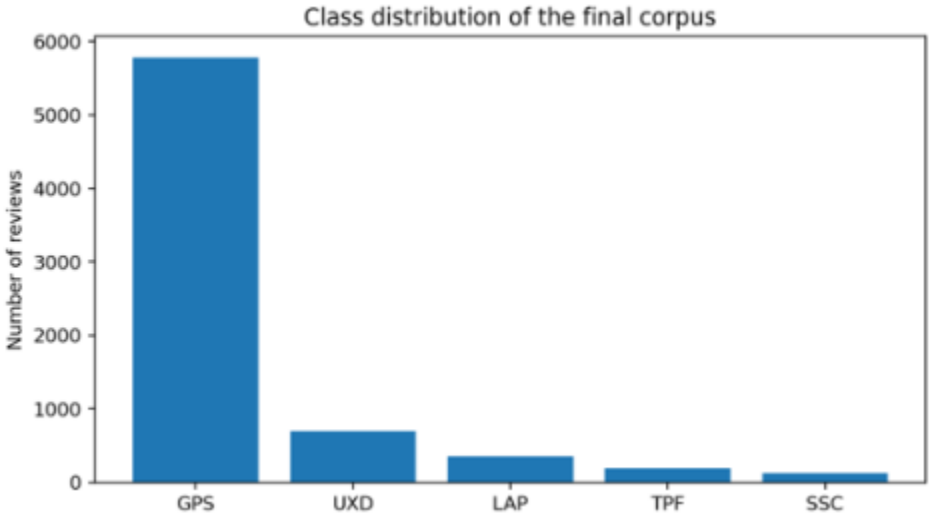

Figure 1

Final class distribution of the 7,148-review corpus

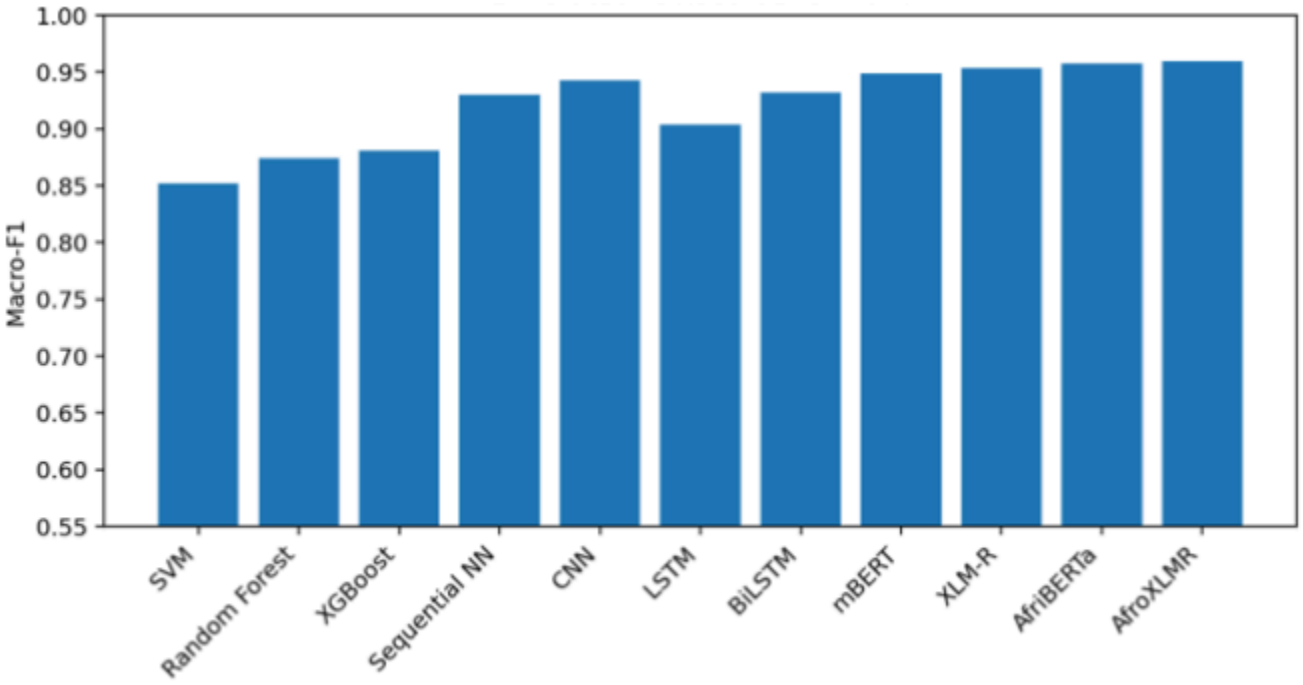

Figure 2

Macro-F1 performance across conventional, neural, and transformer models