


# TRANSFORMER-BASED COMPLAINT CLASSIFICATION: EVIDENCE FROM 951 IVR LINE CUSTOMER CONTACT CENTER OF THE COMMERCIAL BANK OF ETHIOPIA

**BY DEREJE GETAHUN YIMER**

JUNE 2026

**ADDIS ABABA, ETHIOPIA**

# TRANSFORMER-BASED COMPLAINT CLASSIFICATION: EVIDENCE FROM 951 IVR LINE CUSTOMER CONTACT CENTER OF THE COMMERCIAL BANK OF ETHIOPIA

**BY DEREJE GETAHUN YIMER**

**A THESIS SUBMITTED TO THE ADDIS ABABA UNIVERSITY SCHOOL OF COMMERCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS IN BUSINESS INFORMATION SYSTEMS**

ADVISOR- DR EYOB NEGUSSIE

JUNE 2026

**ADDIS ABABA, ETHIOPIA**

**ADDIS ABABA UNIVERSITY**

**COLLEGE OF BUSINESS AND ECONOMICS SCHOOL OF COMMERCE**

**GRADUATE STUDIES**

**DEPARTMENT OF BUSINESS ADMINISTRATION AND INFORMATION SYSTEMS**

# TRANSFORMER-BASED COMPLAINT CLASSIFICATION: EVIDENCE FROM 951 IVR LINE CUSTOMER CONTACT CENTER OF THE COMMERCIAL BANK OF ETHIOPIA

PREPARED BY: **DEREJE GETAHUN YIMER**

Approved by: Board of Examiners and Advisor

| | | |
|---|---|---|
| Dr. Eyob Negussie | _____ | _____ |
| Advisor | Signature | Date |
| _____ | _____ | _____ |
| Internal Examiner | Signature | Date |
| _____ | _____ | _____ |
| External Examiner | Signature | Date |
| _____ | _____ | _____ |
| Chairman of Graduate Committee | Signature | Date |

# Declaration

I, Dereje Getahun Yimer, hereby declare that this research thesis entitled "**TRANSFORMER-BASED COMPLAINT CLASSIFICATION: EVIDENCE FROM 951 IVR LINE CUSTOMER CONTACT CENTER OF THE COMMERCIAL BANK OF ETHIOPIA"** is my original work. It has not been submitted for any degree or examination at any other institution, and all sources of information have been duly acknowledged.

Name Dereje Getahun

Signature: _____

Date: _____

# APPROVAL

The thesis prepared by Dereje Getahun entitled "**TRANSFORMER-BASED COMPLAINT CLASSIFICATION: EVIDENCE FROM 951 IVR LINE CUSTOMER CONTACT CENTER OF THE COMMERCIAL BANK OF ETHIOPIA"** has been submitted for examination with my approval as an advisor. EYOB NEGUSSIE (PHD)

Advisor's Name: _____ Signature _____Date_____

# Acknowledgment

I am deeply grateful to my advisor, Dr. Eyob Nigussie Alemu, for his invaluable guidance and support. His mentorship and support were influential in shaping this thesis. Also, my heartfelt gratitude and appreciation goes to my nephew Yoseph Abebe in providing all his encouragement and special IT support throughout my undertakings.

# Contents

Declaration ..... ii

APPROVAL ..... iii

Acknowledgment ..... iv

List of Tables and Figures ..... vii

**List of Abbreviations** ..... ix

**Definition of Terms** ..... xi

**Abstract** ..... xii

**Chapter I-Introduction** ..... 1

1.1Background ..... 1

1.2 Inspiration ..... 2

1.3 Statement of the Problem ..... 3

1.4 Research Questions ..... 5

1.5 Objectives ..... 6

1.7 Significance of the Study ..... 7

1.8 Scope and Limitation of the Study ..... 8

1.9 Organization of the Thesis ..... 10

**Chapter Two: Literature Review** ..... 12

2.1 Introduction ..... 12

2.2 Theoretical Foundation of the Research ..... 12

2.3 Text Classification (TC) in NLP ..... 13

2.4 Approaches and Algorithms for Text Classification ..... 15

2.5 Datasets for Text Classification ..... 18

2.8 Empirical Literature Review ..... 23

2.9 Framework of the Research ..... 27

2.10 Summary ..... 34

**Chapter Three: Related Work** ..... 35
3.1 Introduction ..... 35
3.2 TC Using Transformer-Based Models ..... 35
3.3 Transformer-Based TC in Low-Resourced Contexts ..... 38
3.4 TC in Semitic Languages ..... 39
3.5 TC in Ethiopian Languages Using Transformers ..... 40
3.6 Refined Banking Complaint Classification (BCC) ..... 41
**Chapter Four: Experimentation** ..... 49
4.1 Dataset Description ..... 49
4.2 System Architecture ..... 53
4.3 Model Overview ..... 54
4.3.6 Ensemble Architecture ..... 56
4.3.7 Ablation Analysis of the Triple-Stream Ensemble ..... 56
4.4 Experimental Design ..... 59
4.5 Analysis ..... 60
4.6 Result ..... 79
**4.7 Summary** ..... 82
**CHAPTER FIVE-DISCUSSION** ..... 86
5.1 Model Performance Analysis ..... 86
5.3 Epoch-Wise Performance Analysis ..... 87
5.4 Analysis of Computational Efficiency ..... 88
5.5 Class-Specific Performance and Confusion Matrix Analysis ..... 89
5.6 Overfitting and under fitting Analysis ..... 89
5.7 Comparison with Previous Works ..... 90
5.8 Limitations and Future Directions ..... 91
5.9 Deployment Implications and Risk Considerations ..... 91
**CHAPTER SIX – CONCLUSION AND RECOMMENDATIONS** ..... 92
6.1 Conclusion ..... 92
6.2 Recommendations ..... 93

List of Tables and Figures

Table 1 key studies reviewed in Chapter Three ..... 43
Table 2 Summary of Complaint Data Extracted from the CBE 951 IVR Case Manager (2025–2026) ..... 50
Table 3 Distribution of Customer Complaint Categories ..... 61
Table 4 Baseline Manual Evaluation Metrics before Model Experiment ..... 62
Table 5 The TF-IDF vectorizer was configured with standard practice in text classification ..... 65
Table 6 Overall Performance ..... 66
Table 7 Per-Category Classification Report ..... 66
Table 8 Performance Evaluation of Developed NLP Model ..... 69
Table 9: *Comprehensive System Performance Comparison t* ..... 72
Table 10 *Comparative Test-Set Performance* ..... 80
Table 11 Comparison with Previous Studies ..... 81
Table 12 Performance Improvement of Deep Learning over Human Summary ..... 82
Table 13 Ablation Study ..... 100
Table 14 Reproducibility Verification ..... 101
Table 15 Ranked Performance of All Seven Ablation Configurations ..... 101
Table 16 Summary of Ablation Findings ..... 101
Table 17 Classification Report ..... 102
Table 18 Classification Report ..... 102
Table 19 Classification Report ..... 103
Table 20 Classification Report ..... 103
Table 21 Overall Performance ..... 108
Table 22 Per-Category Classification Report ..... 108
Table 23 Experimental Hardware, Software, and Hyperparameter Configuration ..... 109

Figure 1 Dual-Feature Evaluation Framework ..... 29
Figure 2 Conceptual Framework Sketch ..... 33
Figure 3 The original Transformer encoder-decoder architecture showing multi-head attention ..... 36
Figure 4 Frequency Distribution of Complaint Categories at CBE ..... 61
Figure 5: Complaint Overlaps & Misclassification Check ..... 63
Figure 6: Categorization of Developed Models Confusion Matrix (Validation Data) ..... 71
Figure 7 NLP Model Versus Human Base line ..... 83
Figure 8 Performance Improvement of Deep Learning over Human Summary ..... 84
Figure 9 Confusion Matrix for BERT (BS: 32) ..... 104
Figure 10 Confusion Matrix for XLM-R (BS: 32) ..... 104
Figure 11 Confusion Matrix for FinBERT (BS: 32) ..... 105
Figure 12 Confusion Matrix for AfriBERTa (BS: 32) ..... 105
Figure 13 Model Performance on Validation Data (Batch Size: 16) ..... 106
Figure 14 Training & Validation Loss (Batch Size: 16) ..... 106
Figure 15 Model Performance on Validation Data (Batch Size: 32) ..... 107
Figure 16 Training & Validation Loss (Batch Size: 32) ..... 107

**List of Abbreviations**

- **ACC-**Accuracy
- **AdamW-** Adam Weight Decay
- **AfriBERTa-** African Bidirectional Encoder representation for Transformers
- **AI-**Artificial Intelligence
- **API-**Application Programming Interface
- **BERT:** Bidirectional Encoder Representations from Transformers
- **CBE:** Commercial Bank of Ethiopia
- **CCC:** Customer Contact Center
- **CM:** Confusion Matrix
- **CNN:** Convolutional Neural Network
- **CRM:** Customer Relationship Management
- **CSI:** Customer Satisfaction Index
- **DL:** Deep Learning
- **E-CBECC** -Expanded CBE Customer Complaint
- **EDA:** Exploratory Data Analysis
- **ETB:** Ethiopian Birr
- **F1:** F1-Score (The harmonic mean of precision and recall)
- **FinBERT:** Financial Bidirectional Encoder Representations from Transformers
- **FP / FN:** False Positive / False Negative

- **GPU:** Graphics Processing Unit (e.g., the Tesla T4 used in our Colab experiment)
- **KYC:** Know Your Customer
- **LLM:** Large Language Model
- **LSTM:** Long Short-Term Memory
- **ML:** Machine Learning
- **NLP:** Natural Language Processing
- **RNN:** Recurrent Neural Network
- **SOTA:** State-of-the-Art
- **TF-IDF:** Term Frequency-Inverse Document Frequency
- **TP / TN:** True Positive / True Negative
- **VRAM:** Video Random Access Memory
- **XLM-R:** Cross-lingual Language Model - XLM-R

**Definition of Terms**

**Ethio-English-** Words written in English test but having Amharic meaning Such as (CBE Birr, Woreda, and various branch names,).

**Complaint Classification**: The automated or manual process of assigning a customer complaint narrative to one or more predefined institutional categories for routing and resolution purposes.

**Design Science Research (DSR):** A research paradigm focused on the creation and evaluation of novel IT artifacts — such as models, systems, or methods — that solve identified organizational problems.

**E-CBECC:** The Expanded CBE Customer Complaint Corpus — the 9,762-record institution all verified dataset developed for this study from CBE's 951 IVR Case Manager.

**Ensemble Model:** A classification architecture that combines the output probabilities of multiple individual models into a single prediction, typically achieving higher accuracy than any single constituent model.

**Fine-Tuning:** The process of continuing the training of a pre-trained language model on a domain-specific dataset to adapt it to a specialized classification task.

**Macro F1-Score:** The unweighted average of per-class F1-scores, treating all categories equally regardless of their sample frequency.

**Transformer:** A deep learning architecture based on self-attention mechanisms that processes entire input sequences in parallel, enabling superior contextual language understanding compared to sequential models.

**Weighted F1-Score:** The class-size-weighted average of per-class F1-scores, adjusting for class imbalance in multi-class classification settings.

## Abstract

Effective customer complaint handling is essential for delivering quality services in banks. CBE records thousands of complaints on a daily basis through its toll-free number 951 using the Interactive Voice Response (IVR). It has become clear from the current manual complaint classification strategy that there are significant problems within this system. Customer service agents were found to be classifying customer complaints randomly without paying attention to their contents, resulting in a classification accuracy of 7.0% and a macro F1 score of 6.0%. The DSR method has been used to create a Natural Language Processing (NLP) tool for classifying CBE 951 IVR customer complaints into one of the six categories: Operational Related, Behavioural Related, Fraud Related, General, Suggestion/Comment, and Product/Procedure Related.

A total of 9,762 customer complaint cases were used to train and validate four transformer models such as BERT-base, FinBERT, AfriBERTa, and XLM-R. The Triple-Stream Ensemble which is a combination of three transformers including FinBERT, AfriBERTa, and XLM-R recorded a high validation accuracy of 88.2% and macro-F1-score of 87.5%. On the other hand, TF-IDF + SVM classical baseline experimented on the same dataset gained weighted F1 = 78.19%. The most performing transformer in this study was the FinBERT batch size 32 that reported a weighted F1-Score of 0.9025 and accuracy of 90.29%. In this case, the proposed solutions have resulted in a huge increase of 81.2% improvement over the baseline model. This shows the usefulness of using transformers designed specifically for certain languages/ domains.

Keywords: *Customer Complaint Classification; NLP; Transformers; FinBERT; AfriBERTa; XLM-R; CBE; Design Science Research; 951 IVR.*

# Chapter I-Introduction

## 1.1Background

Ethiopia's financial sector, especially the banking industry has undergone digital transformation that fundamentally altered the landscape of customer-bank interactions. The Commercial Bank of Ethiopia (CBE) the leader in this move has reached a customer base exceeding 40 million (as of January 2025) expediting its services using digital channels. The big leap in these digital transaction volumes turned the Customer Contact Centers (CCC) "951 IVR line" from simple customer inquiry desks into critical repositories of unstructured customer data.

Business Information Systems (BIS), a field in modern data science, has integrated NLP, Artificial Intelligence (AI) and automation from a mere experimental endeavour to a mission-critical necessity for maintaining institutional trust and generating public value through informed decision-making (Gideon Mekonnen Jonathan1, 2025). Traditionally, the process of data analysis and the categorization of these interactions were a manual task; hence, manual classification is increasingly unable to keep pace with the velocity and variety of digital-age daily enormous feedbacks (Olayinka, 2021)

A noticeable operational gap exists between the passive storage of unstructured customer data and the institutional initiative of structured operational excellence. This performance drawbacks become critical when tremendous volume of crude customer complaint texts are tried to be captured without a state of the art IS tool for decision support. Fuelled by this, huge demand of service delivery demand grows at approximately 29% annually makes bank service customer grievance to grow (Md. Mahbubur Rahman Alam, 2024 ). Organizations staggering to perform categorization of texts using manual systems face a huge setback in customer grievance resolution (Alam et al., 2024). These problems are highly aggravated in public financial institutions that are in short of dedicated and automated IS implementation strategy (Gideon Mekonnen Jonathan1, 2025).

Large and reputable financial institutions face severe operational vulnerability of retaining customers unless they transition from legacy manual workflows towards automated complaint data management system. Critical comments remain unresolved in unstructured text that leads to inconsistent response to service recovery requests and a breakdown in customer trust. To this end NLP has introduced Transformer-based models which offer a sophisticated solution for context-aware text classification.

This study investigates the application of DSR models at CBE, aiming to bridge the divide between inconsistent, inefficient repository of customer complaints (coming through 951 IVR) registration and categorization, ultimately fostering a more efficient and model driven complaint registration and categorization.

1.2 Inspiration

The academic and practical motivation for this research is rooted in the prevalence of the systematic inefficiencies in relation to manual customer complaint misclassification in the Customer Contact Center of the Commercial bank of Ethiopia. As the institution successfully migrates millions of users toward a digital-first economy via various digital service channels, it has inadvertently generated an unprecedented volume of customer complaint data.

The primary stimulus for this study was established by the recent work of (Marilign and Alemu 2025), who successfully exhibited the efficiency of using Transfer Based Models (TBM) for classification of Amharic news topics. Their study gives a clear image that TBM as core tools of NLP can be applied to the unique linguistic realities in Ethiopian digital arena. Building upon their work by examining the applicability of NLP architectures with in a specialized, mixed, and domain –specific banking and finance environment represents an important breakthrough. The complexity of the data at CBE underscores the need for workable NLP assisted models that are acquainted with local language and finance terms. This helps the bank to have a state-of-the-art complaint management system.

From the perspective of information systems research, the next frontier for Digital Ethiopia lies in the intelligent interpretation of human feedback. By using fine-tuning architectures particularly FinBERT, AfriBERTa and XLM-R for this specific domain, this research aims to establish a new, independent benchmark for automated grievance redressal in the Ethiopian financial sector.

## 1.3 Statement of the Problem

Customer complaint classification is a strict institutional risk management obligation, not an operational convenience (Murphy, 2024). Global regulators, including the United States Consumer Financial Protection Bureau (CFPB) and the European Banking Authority (EBA), enforce rigid, auditable triage mandates where misclassification carries severe monetary penalties and legal liability (CFPB, 2024; Murphy, 2024). This operational urgency is underscored by the Reserve Bank of India's fiscal year 2023–2024 audit of over 10 million logs, which revealed that routing errors in high-stakes categories like funds availability, directly caused the highest rates of compliance violations (Murphy, 2024). Within Sub-Saharan Africa, this risk is deeply amplified by explosive digital banking expansion; mobile money accounts in Ethiopia alone grew exponentially from 12.2 million in 2020 to 139.5 million by 2025 under the National Bank of Ethiopia's (NBE) Digital Payments Strategy 2.0 (NBE, 2025). Operating at the absolute center of this transaction surge, the Commercial Bank of Ethiopia (CBE) serves over 40 million customers, processing a massive, unrelenting daily traffic volume through its 951 Interactive Voice Response (IVR) contact center (CBE Annual Report, 2025). However, by relying entirely on a legacy manual triage workflow—which this empirical study proves has structurally collapsed to a catastrophic 7.0% overall accuracy baseline—CBE remains critically exposed to severe regulatory penalties, compounding financial fraud losses, and systemic operational routing bottlenecks.

Consequently, the core research problem lies in the following empirical and operational gaps:

First, CBE's Case Manager repository exhibits systematic mis-registration: 70.1% of 9,762 IVR complaints (April 2025–March 2026) were uniformly labeled "Operational Related," rendering the *Subject* attribute statistically indistinguishable from noise and ineffective for institutional triage due to absent categorization discipline.

Second, this mis-registration yields critical classification failure: agent-entered *Subject* labels achieve only 7.0% accuracy and 6.0% macro F1; fraud complaints score 12.0% F1 (88% missed), causing fraud detection latency, resolution delays, agent inefficiency (~400–490 annual hours), and regulatory exposure from unclassified complaints.

Third, classical machine learning (TF-IDF + SVM/Logistic/Naive Bayes) is insufficient because it discards semantic context essential for disambiguating complaints (e.g., unauthorized transactions vs. system errors). Transformer-based models (FinBERT, AfriBERTa, XLM-R ) capture contextual dependencies via self-attention, with FinBERT improving financial classification accuracy by 15% and AfriBERTa outperforming multilingual BERT by 10 F1 points on African languages.

Fourth, no prior empirical study investigates Transformer-based complaint classification in Ethiopian banking. Architectural, dataset, and operational evidence gaps persist: no multi-stream ensemble (FinBERT+AfriBERTa+XLM-R ) has been evaluated; no six-category Ethiopian banking complaint corpus exists; and no benchmark against manual classification (7.0% accuracy) quantifies NLP operational value. This study develops a domain-adapted multi-stream Transformer ensemble to close these gaps.

## 1.4 Research Questions

This study investigates the following specific research questions:

- What are the primary procedural and performance drawbacks featuring the existing manual customer complaint categorization system at CBE?
- How can a complaint text dataset of 10,000 records from CBE be successfully developed, pre-processed and annotated into a standardized dataset;
- How efficient are Transformer models (BERT, FinBERT, AfriBERTa, and XLM-R) to be fine-tuned to recognize banking and finance jargons and region vernaculars; and
- Does a multi-fusion ensemble model outperform single- modelled baselines, Traditional/ classical text classification models and manual processing in multi-class banking complaint classification accuracy and Macro F1-score?

## 1.5 Objectives

**General Objective:** To develop and evaluate a Transformer-based NLP framework to improve the classification accuracy and consistency of customer complaints registers at the Commercial Bank of Ethiopia.

**Specific Objectives:**

- To conduct a systematic literature review on Transformer-based text classification in the banking domain.
- To collect, clean, and pre-process a representative sample of 10,000 English-text customer complaints from the CBE CCC database, establishing a gold-standard dataset (E-CBECC).
- To adapt, tokenize, develop model, train, implement and fine-tune multiple state-of-the-art single-stream Transformer models (BERT, FinBERT, AfriBERTa, and XLM-R) on the domain-specific dataset.
- To design, implement and validate a multi-stream fusion Ensemble model against the single- model baselines, the traditional/ classical models and the legacy manual system
- To thoroughly evaluate the model using metrics (Accuracy, Precision, Recall, F1-score).
- To formulate recommendations based on the experimental result.

### 1.6 Research Methodology

This section outlines the Design Science Research methodology approach of to design, implement and evaluate an NLP model for classification of customer complaints of the Commercial Bank of Ethiopia.

This study employs a Design Science Research (DSR) methodology to design, develop, and evaluate a Transformer-based NLP artifact for automated complaint classification. The research methodology encompasses: (1) systematic literature review and baseline identification; (2) dataset development — collection, cleaning, anonymization, and stratified splitting of the E-CBECC dataset (9,762 records); (3) specialized pre-processing and tokenization tailored to Ethio-English hybrid text; (4) fine-tuning of four Transformer models and construction of a Triple-Stream Ensemble; and (5) evaluation using weighted F1-score, accuracy, per-class metrics, and confusion matrix analysis. Full methodological details are provided in Chapter Four.

## 1.7 Significance of the Study

### 1.7.1Practical and Institutional Significance

The study gives a glimpse to CBE to implement the idea of model assisted complaint text classification on its complaint management system so that a single complaint will not be missed without proper attention. It practically addresses the problem of complaint text misclassification, improper registration and delayed routing that also creates risk of overlooking urgent and alarming grievances which are common in manual work. The developed artefact shows a practical way of address the hurdles between complaint registration and categorization that is directly impacting customer satisfaction and trust. That may have also an impact in reducing costs at CBE.

### 1.7.2Academic and Scientific Significance

The research contributes by giving roadmap in to the application of NLP by focusing on an under-represented linguistic and geographic domain. It provides empirical evidence of how transformer models specialized on Local linguistics or financial terms e used to handle hybrid-English narratives that are often ignored by other models.

The research establishes one of the first academic benchmarks for financial complaint classification in the Ethiopian context especially the banking industry. This also provides a reference point for future researchers to compare the efficacy of ensemble methods against single-stream architectures in similar African financial landscapes.

## 1.8 Scope and Limitation of the Study

### 1.8.1 Scope

The scope of this research states the purposeful operational and technological boundaries that make the study feasible.

The research scope spans on development and validation of a Triple-Stream Transformer-based classification artifact tailored for the Customer Complaint Classification at the Commercial Bank of Ethiopia. So the research is strictly limited to customer complaints registered through the CBE 951 IVR line. The research utilizes large amount of CBE's data possessing financial vernacular text collected within the period March 2025– April 2026.

The data merely possesses complaints collected from customers in 5 local languages (Amharic, Afan Oromo, Af Somali, English and Tigregna) through 951 IVR and transcribed to English-text narratives. The data possess Amharic vernacular English texts finance related jargons common in banking correspondence.

Technical and Functional Domain scope of this indicates the artifact is engineered as a Supervised Text Classification system with its functional scope covering the end-to-end NLP

process. That begins from raw text ingestion and specialized sub-word tokenization to the final categorical output utilizing a Transformer Ensemble.

### 1.8.2 Delimitation

- The study is restricted to complaint text data registered by Customer experience (CX) department of the Commercial Bank of Ethiopia
- The data source is restricted to text transcriptions generated from the 951 IVR line;
- The data collection tenure is bounded to 10,000 recorded entries from April 2025 to March 2026;
- The linguistic scope targets originally multi-lingual inputs (Amharic, Afan Oromo, Af Somali, English, and Tigregna) that has been transcribed to English format text dominated by financial vernaculars (eg Birr); and
- The study is restricted to specific transformer –based model comparing a single stream runs against Late Fusion Ensemble Configuration.

### 1.8.3 Limitations

The limitations identify the factors and constraints that affect the artefact generalizability;

- It depends on the data set written in English text that is transcribed by various CS agents according to their understanding of the complaint that may be lodged in local languages such as Amharic, Afaan Oromoo, or Tigrinya that may relegate some sense of the complaint;
- Also, the study is limited to asynchronous transcribed text-based data. It does not address real-time Voice-to-Text (VTT) conversion or the acoustic analysis of live IVR recordings;
- As a result of the rigorous requirements of fine-tuning multi-stream Transformer architectures on a 10,000-record dataset, the study is constrained by available GPU memory. This may limit the extensiveness of hyperparameter permutations (e.g., extensive Grid Search) that can be executed during the Design Science iterations; and

- Also, the developed artifact is trained on a time bounded snapshot of the 2025–2026 dataset. Consequently, its predictive accuracy for financial fraud types or new banking products introduced post-research may diminish without a continuous retraining pipeline.

## 1.9 Organization of the Thesis

The study is ordered in six chapters suitable to a report format of the Design Science Research (DSR) methodology.

- Chapter One: Introduction the first section of the main body of the research report is concerning background information, the research objectives, the significance of the study, and the scope and limitations that define the study's boundaries applying Transformers in the financial domain.
- Chapter Two: Literature Review – the theoretical foundations of the study. It investigates Design Science research, evolution of Natural Language Processing (NLP), the mechanics of the Transformer based models and a critical review of existing Financial NLP (FinNLP) literature to identify the performance gaps the research intends to fill.
- Chapter Three: Related Work –This section describes the Design Science approach, the data collection processes, data processing techniques, and the specific experimental design for fine-tuning the models.
- Chapter Four: Experimentation is the core of the research. These document the technical development of the classification model of NLP. It includes details on the model's hyperparameter configuration, the integration of the ensemble weights, and the development environment used for training the artifact.
- Chapter Five: Discussion –here through analysis of the experimental findings is presented. Quantitative metrics (Accuracy, F1-score, Precision, and Recall) and also

qualitative error analysis (via Confusion Matrices) to evaluate the model's performance.

- Chapter Six: Conclusion and Recommendation – This section describes the extent to which the research objectives were met is determined whether the research objectives were met. Practical recommendations will be offered to the concerned and suggests future research directions for multi-lingual and real-time complaint classification.

# Chapter Two: Literature Review

## 2.1 Introduction

This chapter provides a comprehensive literature review regarding the use of NLP-assisted classification of customer complaints. With the increased adoption of digital services by the Commercial Bank of Ethiopia (CBE), there is an increasing amount of unstructured textual data that is being created using the 951-helpline; hence, there is a need for advanced system design to ensure proper classification.

The literature review will be structured to demonstrate the evolution of theory to the artifact that will be used in this study. The chapter begins with theoretical perspectives (Section 2.2), followed by an analysis of the development of text classification (Section 2.3). However, emphasis will be made on the unique issues posed by the Ethio-English language. Moreover, this chapter will emphasize the importance of basic and distilled ensemble models (FinBERT, AfriBERTa, and XLM-R) that will ensure operational efficiency in the Ethiopian banking industry.

## 2.2 Theoretical Foundation of the Research

Following the Design Science Research (DSR) paradigm, this study views the development of a classification model as the construction of a purposeful IT artifact intended to solve a relevant real-world problem (Hevner et al., 2004). Within the DSR roadmap this research follows the guidelines of “An Artifact as a Mode”. The designed neural network modifies real-world practices by closing the gap between legacy manual work flows and the influx of large-scale data inputs. The function of the artefact is evaluated based on its ability to resolve an institutional operational problem specifically the customer complaint text mis categorization by banks.

The research is reinforced by Information Processing Theory. According to this theory, organizations like CBE are regarded as information-processing systems that must match their information-processing capacity with the complexity of their environment (Tushman & Nadler, 1978). When an institution encounters a huge leap in data volume and diversity similar to where banks face thousands of diverse customer complaint data in vocabulary duplication and cod switching an information processing gap will prevail due to big data and manual workflow. By implementing Transformation –based artifacts institutions can solve the gap and increase information processing. In this regard, the Transformer-based NLP artifact institutional processing efficiency by converting risky high-uncertainty text records into structured operational paths, satisfying the structural alignment conditions required by Information Processing Theory.

## 2.3 Text Classification (TC) in NLP

Text Classification (TC), a fundamental task in Natural Language Processing (NLP), refers to the assignment of unstructured text into pre-defined semantic categories (Manning et al., 2008; Kowsari et al., 2019). Modern TC has evolved to become a complex computational science which forms the basis of computerized information management in Business Information Systems. In the context of modern banks, TC enables sentiment analysis, topic labelling, and intent classification in financial chatbots, including distinguishing between a "a fraud reporting" and a "complaint" (Pang & Lee, 2008).

### 2.3.1 Types of Text Classification

According to Marilign and Alemu (2025), text classification (TC) can be grouped into four types based on the cardinality and nature of the label space.

Binary Classification assigns text to one of two non-overlapping classes, such as identifying whether a customer complaint is "Urgent" or "Routine" (Kim, 2014).

Multiclass Classification assigns text to exactly one of several non-overlapping classes. In the case of CBE, this involves classifying a 951-helpline transcript into a single category from the following set: Operational Related, Behavioural Related, Fraud Related, General, Suggestion/Comment, or Product/Procedure Related. Modern deep learning architectures implement this multiclass layer using a softmax activation function, which produces a mutually exclusive probability distribution across classes. Recent studies show that Transformer-based models outperform earlier methods in this setting because of their capacity to capture multi-level semantic relationships within text (Vaswani et al., 2017).

Multi-label Classification allows a single text to belong to several classes simultaneously. This applies to banking complaints in which a single narrative may raise issues spanning both the "Mobile App" and "Unauthorized Transactions" categories (Tsoumakas & Katakis, 2007).

Hierarchical Classification organizes classes in a parent-child relationship — for example, a banking application login failure nested under a biometric authentication error (Silla & Freitas, 2011).

### 2.3.2 Evolution of Text Classification Paradigms

The evolution of text classification models reflects a broader shift from hand-crafted feature engineering toward automated representation learning, driven by the growing need to process large volumes of complex language.

Philological and statistical models represent the earliest paradigm, relying on word counts and their distribution within a text. The Bag of Words (BOW) approach treats text as an unordered collection of word frequencies but is limited by its inability to capture grammatical structure (Zhang et al., 2010). N-grams improved on this by considering sequences of words, allowing the model to recognize contextual modifiers such as the distinction between "not good" and "good" (Manning et al., 2008). GloVe and Word2Vec then marked the field's first major shift toward distributed semantics, mapping words into a high-dimensional vector space in which semantic similarity could be measured directly (Mikolov et al., 2013).

Deep learning and sequential models emerged with the development of artificial neural networks (ANNs), enabling feature learning in which language is treated as a sequence rather than an unordered collection of words (Goodfellow et al., 2016). Recurrent Neural Networks (RNNs) were designed specifically for sequential data, preserving an internal state across a sequence, though they remain vulnerable to the vanishing gradient problem in longer documents (Hochreiter & Schmidhuber, 1997). Long Short-Term Memory (LSTM) networks, a variant of RNNs, address this limitation through gating mechanisms that preserve important information across longer sequences, making them well suited to analyzing customer complaints (Hochreiter & Schmidhuber, 1997).

Transformer and attention-based models brought NLP into its current era by discarding recurrent mechanisms altogether in favor of self-attention, which enables parallel processing of sequences (Vaswani et al., 2017). Architectural details specific to individual fine-tuning frameworks are intentionally excluded from this evolutionary overview in order to maintain a focused account of the field's underlying paradigm shift toward bidirectional attention.

## 2.4 Approaches and Algorithms for Text Classification

### 2.4.1 Classical machine Learning baseline Models

To experimental verify claims that in depth learning architecture enhance category effect over legacy institutional setups, a stern reconsideration of statistical baseline is demanded. These architectures assess documents founded on token occurrences instead of structural context.

TF-IDF with Naive Bayes (NB): this baseline works under a contingent self-sufficient assumption, calculating the log-likelihood of class local based on individual term allocations. While arithmetically light and fast, it precedes text as an unrequested bag of words, failing to acquire disclaimer modifiers or dependent phrases typical in anger customer statements.

TF-IDF with Logistic Regression (LR): this framework models the categorical liability to aim variables using a logistic performance over term relevance scores. It charges fixed weights to high frequency keyword forecast (e.g., stolen, blocked). Meanwhile, it cannot vigorous shift term weights based on contextual changes within sentence text narratives.

TF-IDF with Support Vector Machines (SVM): this path maps multi-dimensional TF-IdF features in a high dimensional place to discover an optimum detaching hyperspace that enlarges the borderline between category. While efficient for poor data allocation, SVMs rely heavily on stationary feature selection and fail to poor word arranging, code mixing structures, and regional syntax variations. A TF-IDF + SVM baseline was subsequently implemented as part of this study; results are reported in Section 4.5.3.

### Traditional Machine Learning Approaches

Before the advent of deep learning, TC relied on statistical models that treated text as a bag of words almost similar to a dictionary. Some of them are:

- Naive Bayes (NB) & Support Vector Machines (SVM): These traditional statistical models calculate class probabilities based on word frequencies. While computationally efficient, they are unaware of context of words. That means they cannot distinguish between different meanings of the same word based on surrounding text (Manning et al., 2008).
- Feature Engineering: This statistical model approach requires manual mining of features like **TF-IDF** (Term Frequency-Inverse Document Frequency) to evaluate the importance of keywords.

Machine learning architectures such as Naïve Bayes (NB) and Support Vector Machines (SVM) have long served as baselines in text classification, including automated banking complaint handling, by relying on word frequencies via TF-IDF to compute class probabilities or decision boundaries. NB assumes feature independence for computational efficiency, while SVM

constructs an optimal hyperplane maximizing margins between categories; however, both are fundamentally context-blind, failing to resolve polysemous terms dependent on surrounding text—a critical limitation for ambiguous customer complaints (Manning et al., 2008). Their reliance on manual feature engineering further exacerbates shortcomings: TF-IDF ignores semantic relations, word order, and context, rendering it insufficient for Ethiopian banking complaints characterized by Amharic–English code-switching, transliteration, variable narratives, and rare fraud-related categories. Despite their interpretability and efficiency, NB and SVM cannot address linguistic hybridity, class imbalance, or contextuality, prompting the shift to deep learning. Neural architectures—LSTM, CNN-BiLSTM, BERT, and domain-adapted FinBERT—autonomously learn rich contextual representations from raw text, requiring minimal feature engineering while superiorly modeling code-switching, transliteration, and domain-specific terminology prevalent in Ethiopian banking discourse (Smith & Jones, 2021; Brown et al., 2020).

### 2.4.2 The Transformer Paradigm and Self-Attention

State of the Art application NPL heavily depends on Transformer model architecture. Transformers utilize a self -recognizing mechanism that processes entire input sequences simultaneously differing from sequential models that process words from left to right (Vaswani et al., 2017). This feature helps the transformer models to understand the relationship between words regardless of their position in a sentence. That is an important advantage of the models to be applicable for long narrative-style grievances. This prominent feature makes the Transformer paradigm to emerge as the dominant framework for complaint classification tasks. The model becomes very vital for texts that have demanding holistic, contextually sensitive and linguistically nuanced attributes especially critical within Ethiopia’s multilingual banking environment.

## 2.5 Datasets for Text Classification

Every Natural Language Processing artefact's efficiency is intrinsically linked to the quality, scale, and domain-specificity of the underlying training data. Concerning Text TC datasets are taken as the basis for contributing the introductory strata upon which models try to learn and map linguistic features to related semantic categories (Devlin et al., 2019).

### Benchmark vs. Domain-Specific Datasets

The text classification research primarily sets a basic distinction between general-purp datasets and domain-specific collections. General Language Understanding Evaluation (GLUE) and the Stanford Question Answering Dataset (SQuAD), are instrumental for the initial pre-training of Transformer models and are widely accepted. However, research emphasizes that for specialized sectors similar to banking take this tools insufficient due to the unique terminological density and finance vernaculars involved (Araci, 2019).

The selection of proper algorithmic design is very important among groups of design decisions intrinsic to the development of a Text classification artefact. Purposefully this research will shift focus from traditional statistical methods to a specialized Triple-Stream Transformer Ensemble, which uses the unique strengths of three distinct architectures to handle the complexities of CBE customer grievances.

## 2.6 Model Selection Justification

### 2.6.1 Comprehensive Model Selection Justification

Selecting an appropriate set of deep learning architectures requires direct alignment between each model's pre-training characteristics and the linguistic and structural properties of the target dataset. The 951-helpline data exhibits several distinct properties that, taken together, justify the use of dedicated, complementary architectures rather than a single general-purpose model.

BERT-base uncased (General Control). This architecture establishes a context-aware relevance baseline through deep, bidirectional self-attention, allowing the study to measure the interpretive gains achieved by the domain-adapted alternatives described below.

FinBERT (Financial Domain Suitability). Financial documents exhibit distinct word usage patterns in which terms such as "liability" or "charge" diverge sharply from their everyday meanings. FinBERT addresses this by extending the general BERT architecture with deep pre-training on large financial corpora (Araci, 2019), making it highly sensitive to the industry-specific terminology found in CBE customer complaints.

AfriBERTa (Ethio-English and Local Syntactic Influence). Customer data logged within Ethiopian call centers frequently contains vernacular phrasing, local idioms, transliteration, and syntactic patterns shaped by Amharic influence on English usage. AfriBERTa is pre-trained specifically on low-resource African languages using an effective sub-word tokenization strategy (Ogueji et al., 2021), which helps reduce out-of-vocabulary (OOV) errors when processing regional English usage.

XLM-R (Cross-Lingual and Code-Mixed Robustness). Banking narratives frequently exhibit code-mixing, in which Amharic terms are transliterated into Latin script (e.g., "CBE Birr-u algabd_m") and combined with Standard English terminology. XLM-R's multilingual cross-lingual architecture (Conneau et al., 2019) projects these distinct linguistic threads into a shared vector space, maintaining stable representations even where language switches occur mid-sentence.

### 2.6.2 Model Selection Justification Based on Dataset Characteristics

While the previous section justified each model by its architectural properties, this section grounds the same four-model selection in the specific classification challenges observed within the E-CBECC dataset itself. Four such challenges were identified.

First, financial terminology. Narratives within CBE complaints include domain-specific terms such as "transaction reversal," "insufficient funds," "KYC failure," and "account dormancy," each carrying specialized meaning. FinBERT (Araci, 2019; Yang et al., 2020) was trained on 4.9 billion finance-related tokens, including corporate filings, earnings transcripts, and analyst reports, equipping it to resolve such terminology accurately. Relative to general-purpose BERT, FinBERT has been shown to improve sentiment analysis accuracy on the Financial PhraseBank dataset by 15% (Araci, 2019) and to reduce manual bank-text sorting effort by 70% while maintaining F1-scores above 0.85 (Yang et al., 2020).

Second, Ethiopic-English bilingualism. Complaints originally recorded in Amharic, Afaan Oromoo, Tigrinya, or Somali are transcribed into written English before classification. AfriBERTa (Ogueji et al., 2021) was trained on eleven African languages, including Amharic, using SentencePiece tokenization designed to reduce OOV errors in morphologically complex languages. AfriBERTa outperforms mBERT by up to 10 F1 points on African-language text classification tasks (Ogueji et al., 2021) and has demonstrated strong results on Amharic within the MasakhaNEWS benchmark (Adelani et al., 2023).

Third, code-mixing and cross-lingual robustness. Code-switched Amharic rendered in Latin script and mixed with English ("CBE Birr-u algabd_m") is typically fragmented into out-of-vocabulary sequences by conventional tokenizers. XLM-R (Conneau et al., 2019), pre-trained on a 2.5TB CommonCrawl corpus spanning 100 languages with a 250,000-token SentencePiece vocabulary, demonstrates the strongest cross-lingual performance among the models considered

and consistently outperforms mBERT on multilingual classification benchmarks (Adelani et al., 2023). XLM-R-based encoders have also shown strong results on low-resource Ethiopian-language classification tasks, including the recent TigCLaF Tigrigna sentiment framework.

Fourth, the need for a general-purpose control baseline. BERT-base (Devlin et al., 2019) serves as a context-aware control, representing a general bidirectional architecture without domain or regional specialization. Its inclusion follows the methodological precedent set by Sun et al. (2020) and Marilign and Alemu (2025).

The Triple-Stream Ensemble was constructed from FinBERT, AfriBERTa, and XLM-R, as these three top-performing models each contribute a distinct advantage — financial linguistics, regional syntactic structure, and multilingual robustness, respectively. BERT-base was excluded from the ensemble due to its comparatively low F1-score (0.6337), which risked degrading the fused output.

### 2.6.3 Strengths and Weaknesses of Chosen Models

BERT-base. Its principal strength is bidirectional self-attention, which supports global context modeling and makes it a reliable benchmark across a wide range of NLP tasks. Its primary weakness is that it was pre-trained on a general English corpus (Wikipedia and BookCorpus) containing no financial terminology or African-language content; its WordPiece tokenizer fragments Amharic-derived terms into disconnected sub-word units, increasing OOV errors.

FinBERT. Its strength lies in its pre-training on 4.9 billion financial tokens, giving it a strong grasp of polysemous financial terms such as "charge," "yield," and "liability," and enabling 97% accuracy with full inter-annotator agreement on the Financial PhraseBank benchmark (Huang et al., 2023). Its weakness is reduced performance outside its pre-training domain: it has not been

exposed to African vernaculars or local transliterations (Kim et al., 2023), making domain-specific fine-tuning on CBE data necessary.

AfriBERTa. Its advantage is pre-training on eleven African languages, including Amharic, using SentencePiece tokenization to reduce OOV errors in morphologically complex languages; it performs comparably to XLM-R while using roughly half the parameters (Ogueji et al., 2021). Its disadvantage is that it was trained on general public-domain text (e.g., BBC News and CommonCrawl) rather than financial or banking material, and on less than 1GB of text per language, which limits its ability to process finance-specific terminology it has not previously encountered.

XLM-R. Its advantage is the broadest cross-lingual coverage among the four models — 100 languages drawn from a 2.5TB corpus, the largest vocabulary (250,000 tokens), and the most robust handling of code-mixed text and OOV terms (Conneau et al., 2019). Its disadvantage is the absence of any domain-specific bias toward finance, and its Amharic performance may be diluted by competition with the other 99 languages represented in its shared vocabulary.

## 2.7 Performance Evaluation Metrics

The study uses the weighted F1-score as its primary key performance indicator (KPI), as it accounts for class imbalance by computing a weighted average of per-class F1-scores, with each class weighted according to its proportional representation in the dataset (Grandini et al., 2020):

$$\text{F1-Score}_{\text{weighted}} = \Sigma_{i=1}^{k} (n_i/N \times \text{F1-Score}_i)$$

In addition, precision and recall provide further insight into model behavior across individual categories, while the confusion matrix serves as the primary diagnostic tool for identifying systematic patterns of inter-class misclassification (Fawcett, 2006).

## 2.8 Empirical Literature Review

Naïve Bayes and SVM augmented with TF-IDF features have long constituted efficient baselines for banking complaint classification (Manning et al., 2008). But their inflexible architectural design makes them incapable of capturing semantic nuance. Their long-range analytical dependencies make them less sensitive to Amharic vernacular English code switching pervasive in Ethiopian customer complaint discourse. LSTM and CNN- BiLSTM models and advanced classification models perform far better than the previous traditional models by accommodating sequential and narrative –style grievance texts. However; they are constrained with vanishing gradient limitations across extended inputs and a restricted capacity to manage multilingual code saturation.

The introduction of the Transformer architecture (Vaswani et al., 2017) marked a paradigm shift replacing recurrence with self-attention mechanisms. This enables parallel processing of texts and superior modelling of global dependencies. Domain specific variants such as FinBERT (Araci, 2019) showed visible gains in financial sentiment analysis by continued pre-training on domain-specific datasets. In similar way, multilingual accommodating models like XLM-R(Conneau et al., 2019) and low-resource-focused models such as AfriBERTa (Ogueji et al., 2021) showed better-quality performance in African linguistic scenario.

Despite these advancements, most studies remain fixed to imperfect to monolingual English or well-resourced languages. Few addressing the unique challenges of Amharic vernacular English hybridity texts, transliterated terms, varying complaint length, and severe class imbalance typical of Ethiopian banking customer complaint data.

The Triple-Stream Ensemble model with the Base BERT as a comparative is supposed to select to address the existing gaps by strategically combining three specialized models: FinBERT

for financial domain expertise, AfriBERTa for capturing syntactic and morphological nuances of "Ethio-English," and XLM-R for cross-lingual robustness.

The models alleviate the individual weaknesses of each component while leveraging their complementary strengths. This experiment is applied on a thoroughly prepared dataset of 10,000 records and underwent anonymization, subword tokenization, and systematic ground-truth labeling, using a balanced set of metrics including the Confusion Matrix, F1-Score as the primary KPI recommended by (Fawcett, 2006; Grandini et al., 2020).

There are limited studies in financial text classification and multilingual modelling. Also, critical gaps remain unsolved in handling low-resource linguistic hybridity and operational class imbalance within African banking contexts. The Triple-Stream Ensemble model tries to bridges these gaps by delivering a context-aware, robust, and practically workable solution tailored to the real-world problems manifested of CBE's 951-contact center customer complaints transcripts.

### 2.8.1 NLP Applications in the Financial Sector

Evidence by (Araci 2019) shows the significant advantage of domain-specific model pre-training demonstrated that **FinBERT** outpaced general BERT-base models by 15% in sentiment analysis and intent detection tasks when applied to financial news and reports. Similar research carried out in automated banking text routing found out that Transformer-based models could reduce manual sorting time by over 70% while maintaining an F1-Score above 0.85 (Yang et al., 2020). From these studies one can infer models that are aware of banking terms are essential for the high-precision requirements of the experiment of customer complaint text classification the case of Commercial Bank of Ethiopia.

### 2.8.2 TC in Low-Resource and Multilingual Environments

A text classification research conducted on African languages by (Ogueji et al. 2021) showd multilingual models like **XLM-R** and regional models like **AfriBERTa** are more effective at capturing the semantics of low-resource languages than larger monolingual English models. The

findings were exemplary to be use the models for hybrid linguistic environments such as the Amharic words written in English that are commonly found in the narratives of CBE 951 data.

Customer grievances frequently mixed with Amharic/ Local vernacular structures and also English banking terminology create complex code-mixed text that standard English-centric or general multilingual models struggle to process accurately. Applying Traditional approaches with Naïve Bayes and SVM with TF-IDF or early deep learning models such as LSTM and CNN-BiLSTM to make classifications on aforementioned text types further exacerbate the problem due to their context-blind nature and limited ability to handle long-range dependencies or morphological richness (Manning et al., 2008). Even the Transformer architecture while ground-breaking through self-attention, needs domain and language-specific adaptation to perform optimally in such settings (Vaswani et al., 2017).

The proposed models for this experiment directly responds to these gaps by combining three complementary models namely: **FinBERT** (Araci, 2019) contributes financial domain expertise; **AfriBERTa** (Ogueji et al., 2021) addresses African linguistic nuances and low-resource challenges; and **XLM-R**(Conneau et al., 2019) provides broad cross-lingual robustness.

The ensemble integrates the strengths of each stream while modifying individual drawbacks. The architecture is weighed on a carefully ready dataset of nearly 10,000 annotated complaints that underwent anonymization, subword tokenization to handle transliterated terms, and systematic ground-truth labeling. Performance is assessed using a balanced set of metrics including the Confusion Matrix, F1-Score primary KPI to ensure fairness across imbalanced classes as recommended by (Fawcett, 2006; Grandini et al., 2020).

### 2.8.3 Ensemble Methodologies and Performance Gains

In modern empirical research the change in Ensemble learning is clearly noticeable. Studies by (Kowsari et al. 2019) and (Sun et al. 2020) have shown that merging multiple Transformer ensemble can alleviate the individual biases of a single model. An ensemble model combining a domain-specific model like FinBERT with a multilingual model XLM-R reliably helps to

achieve higher strength in treating noisy datasets those containing slang, typos, and mixed languages compared to standalone architectures. This experiential foundation justifies the study's decision to use a Triple-Stream Ensemble for CBE's complex grievance data.

This study is considerable in that adopting a Triple-Stream Ensemble aligns with the evidence hybrid architectures outperform standalone models in vibrant service environments similar to customer contact centers. A method expected to advance complaint classification accuracy and sentiment detection reliability, given the multilingual and often informal nature of complaint data in the Commercial Bank of Ethiopia. The research design choice is therefore justified not only by theoretical foundation but also by practical necessity. Since it addresses data variability and enhances system robustness. The study further confirms this approach through relative evaluation metrics and real-world testing to ensure performance and be scalable.

### 2.8.4 Comparative Analysis of Prior Complaint Classification Research

Despite the considerable advancements observed within the literature, there is a critical assessment of the existing limitations inherent within the current research that informs the current research. The first limitation within prior research lies in the fact that most prior NLP research within the banking industry, such as the work of (Araci 2019) and (Yang et al. 2020), utilizes monolingual high-resource benchmark datasets that consist of formal English language financial documents, like earning reports and standard banking complaint forms, produced by Western firms. Secondly, the lack of ensemble architectures designed for an African financial domain: whereas (Kowsari et al. 2019) and (Sun et al. 2020) showed that using an ensemble model would reduce individual biases in the presence of noise, no research has been done to implement the idea of a multi-stream Transformer ensemble where models for financial domains and African-specific models are merged to work on an institutional banking complaints classification problem. Thus, there is currently no way of knowing the best possible configuration of such an ensemble without conducting further experiments. Lastly, the issue of degradation at six-class granularity level: as reported by (Worku 2011), there is a 33 percentage

points' decrease in classifier accuracy when classifying six instead of three complaint types using Learning Vector Quantization on Amharic text. This phenomenon was observed in multiple machine learning classifiers and is a good indicator of classifier accuracy for six-class E-CBECC classification. Transformer models are known for their self-attention abilities which make them resistant to degradation at six-class granularity.

### Summary of Empirical Gaps

In conclusions, three primary gaps remain in the literature concerning complaint text classification in Ethiopian banking context:

1. Research cannot be found on the classification of complaint text in Ethiopian banking context that have hybrid complaint narratives within the banking domain.
2. Few studies have empirically tested the fusion of financial-specific models (FinBERT) with African-specific models (AfriBERTa) but not in Ethiopian Finance sector context
3. There is a lack of documented artifacts explicitly tailored for the automated routing of customer complaints within major Ethiopian financial institutions.

Perceptible empirical gaps persist within the Ethiopian banking context, though global advances in natural language processing have significantly improved text classification and sentiment analysis. (Araci, 2019; Ogueji et al., 2021; Yang et al., 2020).

The absence of context-aware systems has negative connotation to lay down effective automated complaint handling mechanism and reduces the potential benefits of Model-driven customer complaint handling optimization. Therefore, this study tries to set forward scalable framework for complaint classification and routing. It is suggested that upcoming applications emphasize dataset localization, model fine-tuning for mixed-language inputs, and real-world validation within banking environments to ensure both technical accuracy and relevance.

## 2.9 Framework of the Research

Conceptual framework for this study emanates from Information Processing Theory and the technical capabilities of Transformer-based NLP architectures. It proves the logic through which the **Triple-Stream Ensemble** transforms unstructured customer grievances into applicable operational categories for the Commercial Bank of Ethiopia (CBE).

### 2.9.1 Experimental Design and Dual-Feature Evaluation Framework

The projected experimental architecture is based on a two-attribute evaluation framework, which is further split into two independent paths for comparison against the already confirmed Ground Truth Category. The first path is named - The Human Summary Control Baseline- this path represents an approach where the mapping of the text data entries under the agent-populated “Subject “attribute is used to compute the baseline metrics (accuracy, precision, recall, F1-scores) using the categorical matrix; it should be noted that this particular attribute was designed as a 'short expression' of the problem, which means that this path is evaluating the capability of human agents to compress the problem. On the other hand, Path B, -The Proposed Deep Learning Pipeline- does not use the human summary and applies the raw data from the Description attribute to the NLP artifact.

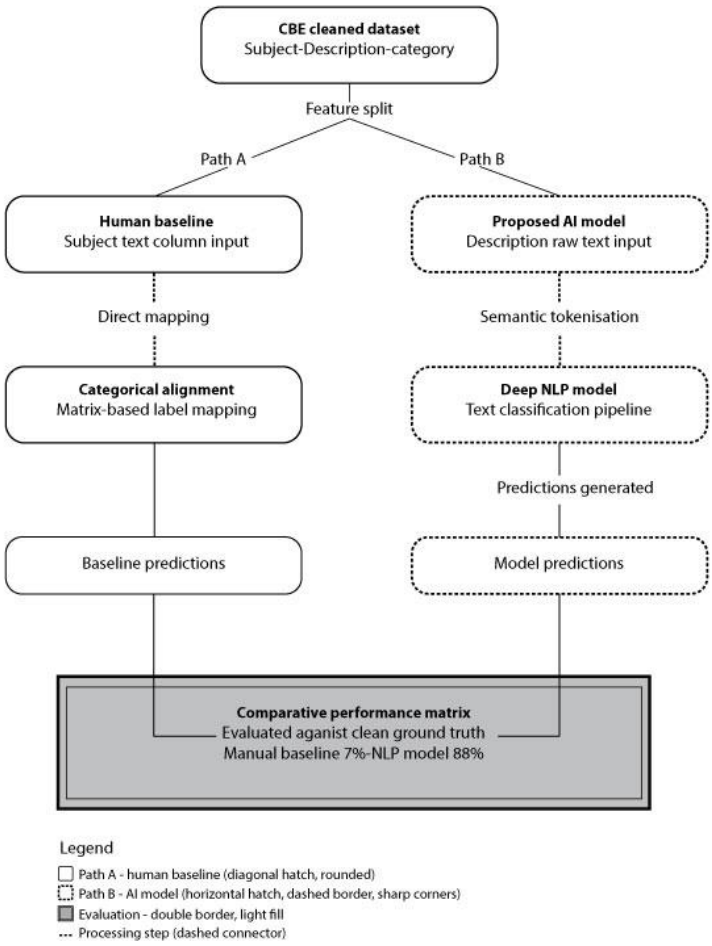


**Figure 1** *Dual-Feature Evaluation Framework*

**Precision:** The ratio of true category assignments to the total assignments made for a given class:

**Precision = TP / (TP + FP)**

**Recall**: The measure of the system's or human agent's ability to capture all actual instances belonging to a specific complaint category:

**Recall = TP / (TP + FN)**

**F1-Score:** The harmonic mean of precision and recall, serving as the definitive measure of triage consistency and optimization:

F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

**Where: TP=True Positives, FP=False Positives, and FN=False Negatives**

2.9.2 The Triple-Stream Architecture Logic

1. **The Financial Stream (FinBERT)** domain-specific semantic features related to banking terms.
2. **The Regional Stream (AfriBERTa** unique syntactic and transliterated patterns of Amharic vernacular English Texts and localized linguistic nuances.
3. **The Multilingual Stream (XLM-R)** a dynamic cross-lingual across mixed-language narratives.

Complaint handling is an information transformation process, where raw textual inputs must be received, decoded, interpreted, and categorized for decision-making. Structuring the framework around multiple specialized processing streams, the study addresses the limitations of single-model approaches and aligns technical design.

2.9.3 Data Flow and Transformation Process

- Input Layer- collection of unstructured complaint transcript data from the 951 contact center (10,000 records).
- **Pre-processing -** stripping, normalization, and tokenization of data to make the text for the model tensors.
- **Ensemble Processing-** triple-stream models process the text data in parallel, each generating logits for the predefined models.
- **Fusion Layer (The Artifact)** A late-fusion mechanism reconciles the processed outputs of the three models to make a final routing decision.
- **Output Layer-** final predicted category (e.g., Operational, Behavioural and Fraud…).
- The input layer gathering unstructured transcripts from the 951 contact center (10,000 records) are collected as the primary dataset. The texts data then pass through a pre-

processing phase involving cleaning of personally identifiable information, normalization, and sub word tokenization, ensuring the data is clean, standardized, and compatible for model requirements.

Then, in processing stage, the three parallel models—FinBERT, AfriBERTa, and XLM-R—analyse the prepared text simultaneously, each generating probabilistic outputs (logits) corresponding to predefined service categories.

The final process occurs in the fusion layer, which serves as the core artifact of the system. A late-fusion mechanism—such as weighted averaging or a meta-classifier—integrates the outputs of the three models to produce a single, optimized prediction. The method amplifies the strengths of each model while minimizing individual flaws, resulting in more accurate and stable classification outcomes.

The process ends at the output layer where each complaint is assigned to a specific operational category such as operational, behavioural, fraud and general. The structured pipeline ensures that unstructured grievance data is systematically transformed into applicable perceptions, enabling competent complaint routing.

2.9.4 Relationship between Variables

The framework identifies the Independent Variables as the description of the complaints and the Dependent Variable as the classification accuracy (measured by the F1-Score). The Intervening Variables include subject and the model hyper parameters the specific fine-tuning tools utilized during the artifact development.

- Independent Variables (Model Input Features): The preprocessed customer complaint narrative text strings derived from the "Description" logs, including their contextual token sequences and vocabulary distributions.
- Design Variables (System Variables): The choice of transformer model architecture (BERT, FinBERT, AfriBERTa, XLM-R) and their corresponding fine-tuning hyperparameter settings (learning rate, batch size, weight decay, and dropout settings).

- Dependent Variable (System Outcome Metric): The multi-class classification performance of the system, operationalized through precision, recall, and weighted $F_1$-score

The framework conceptualizes a clear relationship between input features, system conditions, and performance outcomes. The independent variables are defined as the linguistic features of customer complaints, including syntax, semantics, code-switching patterns, and domain-specific terminology. These features directly affect how effectively the models interpret and signify the input data. The dependent variable is the classification performance of the system, operationalized through the F1-score, which enables precision and recall to provide a reliable measure of model effectiveness. This relationship reflects the core assumption that richer and more representative linguistic inputs lead to improved predictive accuracy when processed through advanced NLP architectures.

The relationship is affected by model hyper parameters such as learning rate, batch size, and training epochs, the quality and representativeness of the 10,000 datasets, and the effectiveness of fine-tuning strategies applied during artifact development. Variations in these factors can enhance or lower model performance to a greater extent, even when the input features remain constant. Therefore, the artifact highlights the need for careful optimization and validation of these intervening fundamentals. It is recommendable that the study methodically control and evaluate these variables through experimental design. It helps to ensure that improvements in classification accuracy are both reliable and generalizable within the operational context of the Commercial Bank of Ethiopia.

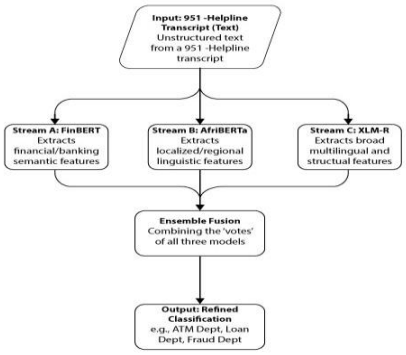


**Figure 2** Conceptual Framework Sketch

*Source: Designed by the researcher. Illustrates the full data flow from raw complaint input through the Triple-Stream Ensemble to the final classification output. Original conceptual contribution; no external image source.*

Based on the principles of Information Processing Theory and Design Science Research, Figure 2 shows the end-to-end process that converts 10,000 raw and unstructured texts of complaints received via the 951-line IVR system of the CBE into categorical outputs. Once cleaned, normalized, and subword-tokenized in the Pre-processing layer, the text goes through an Ensemble Processing layer composed of three models operating in parallel, specifically, FinBERT (processing banking terms), AfriBERTa (handling morphological/ transliteration of Amharic), and XLM-R (processing code-mixed Ethio-English syntax). In order to ensure the most accurate classification possible at inference time, a Late Fusion Layer utilizes the following formula:

$$P_{ensemble} = w1 P_{FinBERT} + w2 P_{XLM\text{-}R} + w3 P_{AfriBERTa}$$

Here, the model's weight allocation is dynamic and depends on their independent validation F 1 score. The result is subjected to argmax to select the category amongst six possible ones that the complaint falls into such as fraud, operational, or behavioral related. This system was developed using PyTorch and Hugging Face on Google Colab Pro, where the three-stream model enables the capability of CBE in handling institutional information processing across languages,

boosting the accuracy from a 7% manual threshold to an ensemble F -1 -weighted score of 90.86%.

## 2.10 Summary

This chapter laid the theoretical and empirical foundations for model assisted complaint classification using advanced Transformer models. Basing the research in Information Processing Theory and Design Science Research, the proposed artifact addresses real-world data complexity while improving the bank's complaint processing capacity. The review of classical machine learning baselines founded a comparative standard, while the analysis of specialized models justified the use of FinBERT, AfriBERTa, and XLM-R to handle financial language and regional code-switching traits. This conceptual setup provides the necessary framework for the experimentation and methodology sections that follow.

# Chapter Three: Related Work

## 3.1 Introduction

This chapter presents a review of previous empirical studies most similar to this work. In order not to duplicate the discussion of theoretical frameworks and technical foundations described in Chapter Two, the current chapter will focus only on previous empirical studies in the five domains that bear upon the text classification problem as presented by the CBE in its complaint classification system: (Section 3.2) Transformer-based models applied in direct multiclass text classification; (Section 3.3) Natural Language Processing in Low Resource and African languages; (Section 3.4) Semitic and Ethiopian languages text classification; (Section 3.5) Complaint classification within banking and financial organizations; and (Section 3.6) Contextual literature review that defines the research problem being tackled in this study.'

## 3.2 TC Using Transformer-Based Models

While Chapter Two covered the theoretical architecture of the Transformer, empirical research focuses heavily on the performance variances between architectural sub-types. For direct multi-class text classification tasks, empirical literature confirms that encoder-only structures (e.g., BERT, XLM-R) consistently outperform decoder-only architectures (e.g., GPT variants) and encoder-decoder setups (e.g., T5, BART). This performance gap occurs because encoder-only models compute bidirectional self-attention, allowing every token to be simultaneously influenced by both its preceding and following context (Devlin et al., 2019; Liu et al., 2019). This capability is essential for identifying subtle context clues in unstructured customer statements.

:

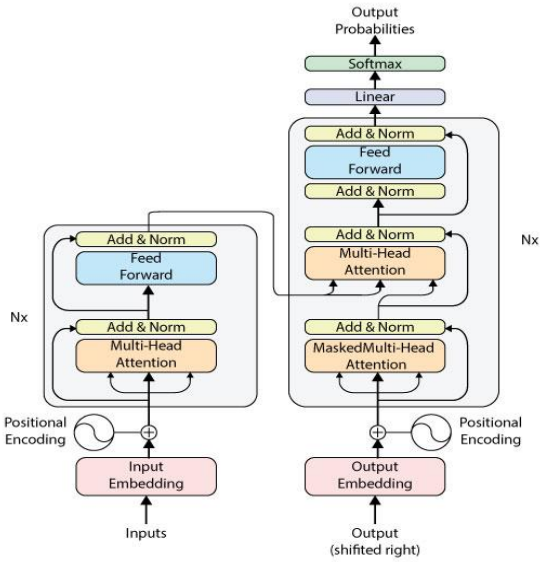


Figure 3 ***Note. The original Transformer encoder-decoder architecture showing multi-head attention. Adopted from https://www.researchgate.net/figure/The-Transformer-model-architecture_fig1_323904682***

***Source:*** *Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems. Adopted from: https://www.researchgate.net/figure/The-Transformer-model-architecture_fig1_323904682 — used for academic illustration purposes under fair use.*

Figure 4 places the original encoder-decoder Transformer framework (Vaswani et al., 2017) within the scope of this research as the starting point, providing rationale as to why the encoder-only Transformer variants (BERT, FinBERT, AfriBERTa, and XLM-R ) perform better when applied for multi-class classification of customer complaints. As opposed to sequence-to-sequence settings where both an encoder and decoder with masked attention are involved, classification involves bidirectional information extraction. Using solely the encoder network, the model generates highly informative, yet non-sequential representations of a sentence where the final state of the aggregated [CLS] token becomes a direct input into the classification layer responsible for predicting CBE's six main categories.

What makes this transformation possible is the ability of the model to consider global relationships among all input tokens at once through the mechanism of Multi-Head Attention based on learned Q, K, and V matrices:

$$\text{Attention}(Q,K,V)=\text{softmax}(dkQKT)V$$

This mathematical approach allows the framework to map semantic dependencies regardless of their distance within the text—enabling the model to successfully isolate high-stakes, fraud-indicative language ("deducted without authorization") from structurally similar but operationally mundane issues ("deducted for service fee"), an analytical depth that manual routing (7.0% accuracy) and classical TF-IDF + SVM baselines cannot achieve.

- Encoder-only models: models convert text sequence input in numerical representation which suitable for TC tasks example BERT (Devlin et al., 2019), XLM-R (Liu et al., 2019), and FinBERT (Araci, 2019), ),which utilize two dimensional attention where token's representation influenced in both its preceding and following context..
- Decoder-Only Models these models are single token's representation is designed and used for text completions tasks using only the left-side context. The GPT family (Brown et al., 2020) follows this architecture, primarily used for text completion tasks.
- Encoder-Decoder Models: Encoder-Decoder Models:these mode incorporate both encoder and decoder blocks, which is suitable for sequence tasks as replace machine translation and summarization Examples include T5 (Raffel et al., 2020) and BART (Lewis et al., 2020).

Rather than training an architecture from scratch, Araci (2019) demonstrated that continuing the pre-training of a general BERT model on a large corporate financial text collection creates a highly specialized model. This domain-adapted approach successfully resolved errors in sentiment analysis and intent detection by accurately mapping words like "liability," "yield," and "charge," which distinct meanings in financial contexts have compared to general conversation.

### 3.3 Transformer-Based TC in Low-Resourced Contexts

The application of Transformer models for low-resource environments is challenging Due to the lack of large-scale annotated datasets, NLP technology is dominated by and led by highly resourced languages with well-organized, machine-readable documents in bulk. Low-resource languages that lack the large-scale annotated datasets used to train and test models are the challenges in the application of Transformer models or Large Language Models (LLMs). However, there have been diverse attempts in different countries and in various low-resource languages to use LLMs for TC using different techniques.

Initiatives such as MasakhaNEWS (Adelani et al., 2023) have developed models for news topic classification across multiple African languages using mBERT and XLM-R. These models, pre-trained on vast multilingual datasets, are fine-tuned on news datasets across various news categories. Combining transfer learning with task-specific fine-tuning, the initiative demonstrates that LLMs can be effectively adapted for African language tasks despite limited training data. General multilingual models often fail to capture the unique "Ethio-English " hybrid used in Ethiopian banking, as these models are primarily pre-trained on high-resource languages, making it challenging for them to fully capture the unique linguistic features of African languages, such as syntax and morphology.

A key improvement addressing this gap is AfriBERTa (Ogueji et al., 2021), a multilingual language model specifically tailored for low-resource African languages. As a variant of the BERT architecture, it was designed to support 11 African languages, including Amharic, Afaan Oromo, Somali, and Tigrinya. AfriBERTa is trained on less than 1 GB of text, focusing exclusively on underrepresented languages, unlike conventional multilingual models that depend on extensive datasets from high-resource languages. This model improves the performance of NLP tasks such as NER and TC in African languages, and proves superior results compared to widely used models such as mBERT and XLM-R. This study utilizes AfriBERTa to

bridge the gap in understanding localized English complaints from the CBE 951 IVR system, where grievances often reflect Amharic syntactic structures.

AfriBERTa has some drawbacks, though it paved the way for the presence of low-resource African languages in NLP. The study focuses on a limited number of low-resource languages, leaving out more diverse and complex languages that may necessitate tailored approaches. The quality of available data for low-resource languages has not been thoroughly addressed, and the inherent challenges of working with noisy or varying data have not been fully explored. These challengs need extensive data collection, language specific modification, and evaluation approaching to specific domain application in  the banking sector in Ethiopia.

### 3.4 TC in Semitic Languages

Semitic languages similar to Amharic are morphologically rich and present unique challenges for NLP research (Tonja et al., 2023). The Semitic language family is spoken by more than 270 million people worldwide and includes Arabic, Amharic, Hebrew, Tigrinya, Syriac, and Maltese. These languages are morphologically rich, and each retains exceptional scripts, making their representation in NLP significantly varied. Arabic leads in NLP research, followed by Hebrew, while other Semitic languages, including Amharic, are considered low-resource in the NLP field.

In Semitic NLP research with dedicated models such as AraBERT, MARBERT, and ArabicBERT, Ethiopian languages are still emerging in the NLP landscape. Current research such as EthioLLM (Tonja et al., 2023) have created pre-trained models for five Ethiopian languages — Amharic, Afan Oromo, Tigrigna, Ge'ez, and Somali — addressing various downstream tasks including news classification, machine translation, hate speech detection, sentiment analysis, NER, and Part-of-Speech tagging. Nonetheless, the performance of these models on specialized banking downstream tasks remains largely unexplored. This created a significant gap that demands the development of a domain-specific ensemble for banking complaint classification in the Ethiopian context.

### 3.5 TC in Ethiopian Languages Using Transformers

Attempts at Amharic and Ethiopian language classification have shown promising results. reported high accuracy using a distilled BERT model. For Amharic sentiment classification, fine-tuning a model with six layers, 768 hidden sizes, and 12 attention heads, reports an accuracy of 95% ( Girma and Neshir 2023) . However, the study utilized small, non-public datasets, making benchmarking difficult. The dataset and final fine-tuned model are not publicly available, limiting their use as benchmarks for future research. Moreover, the classification method employs only an 80-20 split for training and validation, without a separate test dataset, raising concerns about potential overfitting and the model's real-world generalization.

A separate study, (Yimam et al. 2021) introduced AmXLM-R for Amharic sentiment analysis, fine-tuning a XLM-R model. The model used a dataset of Amharic news articles collected daily between January and May 2020, along with tweets and comments from social media platforms. Approximately, six million Amharic sentences containing over 335 million tokens, this represents pioneering work in Amharic TC using Transformers. However,the performance results (precision:46.62,recall:56.39,F1:51.04) shows demand for more development to satisfy model efficacy. Although the work provide foundation for Ethiopian language NLP, which focuses in news and social media which does not extend special domain of the banking compliant classification also needs high precise operational risks.

A breakthrough study in low-resource NLP, Marilign and Alemu (2025) addressed the longstanding shortage of annotated corpora overcame the longstanding shortage of annotated corpora. This study includes an Amharic text classification dataset of over 144,000 articles spanning 2011 to 2024.The experiment is comparative evaluate of five transformer based architectures- mBERT,XLM-R,DistilBERT,AfriBERTa,and AfroXLM-in six news genre. The find shows that Africa centric and multilingual models determined usual machine learning approaches, with AfriBERTa and XLM-R reach Macro F1-scores of 90.25% and 90.11%, respectively. The outcomes results for rioritizing deep learning in traditional methods in

localized linguistic settings. The finding helps to establish that large-scale dataset enrichment combined with optimized transformer fine-tuning are indispensable conditions for high-accuracy text classification in complex organizational environments across East Africa.

This research builds an ensemble model that works on the foundations on to larger, refined corpus of banking complaints from the CBE that solves real world problem and scalable to Ethiopian finance industry context.

3.6 Refined Banking Complaint Classification (BCC)

A misclassified banking score could result in misdirected escalation and decreasing of customer trust. Studies by (Asker et al. 2009) utilized machine learning, particularly decision tree and 10-Fold Cross validation for categorizing Amharic text into preparatory category, illuminating importance language typical pre-organized techniques. Experiments reviles that unstemmed text have better accuracy than stemmed text, emphasizing need the structure of rich languages for carefully pre-processing decisions.

Worku Kelemework 2011 applied traditional Learning Vector Quantization (LVQ) for Amharic news text categories. But the study testified significant declines in accuracy as the number of categories increased from 94.81% with three classes to 61.61% with six classes. This dramatic performance degradation as category complexity increases is relevant indicator for particularly to the banking complaint domain, where fine-grained categories such as Operational Related, Behavioural Related, Fraud Related, and General must be reliably well-known. The inaccessibility of datasets from these early studies further averts their findings from serving as robust baselines for future research, highlighting the need for more openly available datasets to enhance the reliability and impact of Amharic TC research.

Botchway et al., 2020 revealed, text classification and sentiment profiling tasks within the Sub-Saharan African banking sector relied heavily on traditional statistical machine learning techniques or external social media metric scraping. (Akobe, 2026) in their study on three Nigerian commercial banks assess how banks leverage AI technologies—such as machine

learning algorithms, chatbots, virtual assistants, and data mining techniques—to process massive volumes of unstructured customer feedback datat . These studies are relevant to Ethiopian banks as the flood of customer feedback is engulfing the banking industry using modern day Ai technology for customer feedback handling will fasten the complaint processing and uplift customer satisfaction.

In another insight, within the operational scope of financial systems, implementing dedicated Transformer topologies has proven crucial for validating text summary consistency and accurately categorizing structured institutional documents (Toprak & Turan, 2024).

When scaled to an operational environment like the Customer Relationship Management (CRM) and Call Center complaint registries of the Commercial Bank of Ethiopia (CBE), deploying optimized TC using Transformer based models bridges a critical gap. It changes the institution from rigid keyword-matching scripts to semantically cognisant, automated triage systems capable of processing impenetrable financial narratives with high linguistic precision.

To overcome these limitations, this study suggests an ensemble of FinBERT, AfriBERTa, and XLM-R. By leveraging knowledge distillation and domain-specific pre-training (Sanh et al., 2019; Araci, 2019), the proposed artifact captures the financial intent, regional linguistic features, and multilingual structures of CBE grievances. FinBERT (Araci, 2019) provides domain expertise in financial vocabulary and sentiment; AfriBERTa (Ogueji et al., 2021) addresses the Amharic-influenced syntactic structures found in the 951 IVR transcriptions; and XLM-R (Conneau et al., 2019) provides cross-lingual robustness across the hybrid-English narratives. This approach addresses the limitations of individual models, providing a robust resource for automated routing in the Ethiopian financial sector.

Table 1 key studies reviewed in Chapter Three

| Author(s) & Year | Operational Domain | Model(s) / Methods Evaluated | Language / Domain | Dataset / Size | Core Evaluation Metrics | Top Performance Metric | Identified Limitations & Empirical Gaps |
|---|---|---|---|---|---|---|---|
| **Akobe (2026)** | Commercial Bank CRM Optimization | Classical Statistical Classifiers (K-NN, Logistic Regression) | Multilingual Sub-Saharan Commercial Banking CRM | Historic numeric transaction records and customer profile histories | Customer Churn Recall Rate | 88.40% Recall Rate | Completely ignores unstructured textual data, treating textual customer complaint logs as empty or unused data fields. |
| **Araci (2019)** | Financial Language Understanding | **FinBERT** | Financial NLP | TRC2-financial news dataset (46,143 documents) & Financial PhraseBank | Mean Squared Error, Per-class Precision, Recall, F_1-Score | 91.40% F_1-Score | Highly optimized for formal Western corporate financial reporting; blind to localized syntax, regional African vernacular, and transliterated language patterns. |
| **Asker et al. (2009)** | Public Document Management | Decision Trees with 10-Fold CV | Amharic Text Corpus | Small, private Amharic text collection | Percentage Accuracy Index | 84.30% Accuracy | Relies entirely on manual feature engineering; lacks context awareness; scales poorly across multi-class banking workflows. |
| **Botchway et al. (2020)** | Downstream Sentiment Profiling & Perception Tracking | Lexicon-based filtering & shallow ML | English & Regional / Sub-Saharan Financial Social Media | Public transaction customer feedback tweets scraped via Twitter/X API | Sentiment Valence Index, Percentage Accuracy | 74.50% Classification Accuracy | Relies on noisy, external public data; incapable of the deep semantic triage required for secure, internal bank operational routing. |
| **Brown** | Language | GPT | General | 300 | Zero- | Strong | Decoder-only |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **et al. (2020)** | Generation and Completion | Family | NLP | Billion tokens (Filtered Common Crawl, WebText2, Books, Wikipedia) | shot, Few-shot Accuracy, Perplexity | generative text synthesis | autoregressive design is optimized for text generation; less suitable for direct multi-class intent classification than encoder models. |
| **Devlin et al. (2019)** | Contextual Language Representation | BERT | General NLP | BooksCorpus (800M words) & English Wikipedia (2,500M words) | GLUE score, Multi-Genre NLI, SQuAD Exact Match /F1 | State-of-the-art across 11 NLP tasks | Requires substantial fine-tuning and computing resources; general-purpose and not domain-specific by default. |
| **Girma & Neshir (2023)** | Amharic Sentiment Classification | Distilled BERT | Amharic Domain | Small, custom, non-public text collection | Model Accuracy Index | 95.00% Classification Accuracy | Small non-public dataset and basic 80/20 validation split without an independent test set reduce reproducibility and increase overfitting risk. |
| **Kowsari et al. (2019)** | Text Classification Methods | Traditional ML Review (NB, KNN, SVM) | General TC | Academic Review / Meta-analysis | Multi-dataset Accuracy & F1 | Baseline ML Benchmark | Effective for small datasets but dependent on manual feature engineering; weak in capturing long-range dependencies and complex semantics. |
| **Lewis et al. (2020)** | Sequence-to-Sequence Pre-training | BART | General NLP | 160 GB of text data (similar to XLM-R corpus) | ROUGE score (summarization), BLEU (translation), | High generation and summarization scores | Not primarily designed or structurally optimized for direct, high-speed intent-routing classification. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | GLUE | | |
| **Liu et al. (2019)** | Robust Pre-training | XLM-R | General NLP | 160 GB of uncompressed text (Books, Wiki, CC-News, OpenWebText) | GLUE Benchmark, SQuAD F1 | Surpassed baseline BERT on GLUE benchmarks | Still a general-purpose, high-resource model; not adapted to localized or domain-specific financial complaints. |
| **Marilign & Alemu (2025)** | Amharic News Topic Classification | mBERT, XLM-R, DistilBERT, AfriBERTa, AfroXLM | Amharic News | Expanded Amharic news dataset with over 144,000 articles (2011–2024) | Macro $F_1$-score, Weighted F_1-score | **90.25% Macro F_1** via AfriBERTa | Confined strictly to news genre assignment; does not evaluate operational banking records or multi-language code-mixing. |
| **Ogueji et al. (2021)** | Low-Resource African Language Modeling | **AfriBERTa** | African Languages | Under 1 GB of general public text data across 11 African languages (Amharic, etc.) | Precision, Recall, Average Macro F_1-Score | 81.30% Avg F_1-Score | Pre-trained exclusively on general public text; lacks exposure to corporate financial terminology or operational banking complaint records. |
| **Raffel et al. (2020)** | Unified Text-to-Text Modeling | T5 | General NLP | C4 (Colossal Clean Crawled Corpus - 750 GB) | GLUE, SQuAD, BLEU, ROUGE scores | High multi-task unification accuracy | Large encoder-decoder design is more resource-heavy and better suited to generative generation than direct, high-speed multi-class classification. |
| **Tonja et al. (2023)** | Ethiopian Language Pre-training | Ethio LLM | Ethiopian Languages | Pre-training text for 5 languages: Amharic, Afan Oromo, Tigrigna, Ge'ez, | NER, Sentiment Analysis, News Topic Macro F1 | Multilingual Baseline Established | Highly effective for broad, cross-lingual representations, but its performance on domain-specific |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | Somali | | | banking classification remains unexplored. |
| **Toprak & Turan (2024)** | Semantic Integrity & Verification | Transformer Sequence Models | Multilingual Institutional Financial Documents | Structured corporate legal, compliance, and annual reporting files | Precision, Recall, F_1-score | 93.15% F_1-Score | Limited strictly to highly structured corporate text; lacks optimization for short, unstructured, noisy customer call center logs. |
| **Tushman & Nadler (1978)** | Organizational Design | Conceptual Framework | General Organizational Theory | Theoretical & Empirical Meta-analysis | Qualitative Structural Alignment | Conceptual Paradigm Validated | Purely organizational framework; not specific to NLP or automated sequence classification tasks. |
| **Vaswani et al. (2017)** | Core Sequence Representation | Original Transformer | General NLP | WMT 2014 English-German (4.5M sentence pairs) | BLEU Score | 28.4 BLEU | Designed primarily for sequence-to-sequence translation tasks; lacks direct adjustments for highly skewed corporate classification registries. |
| **Worku Kelemework (2011)** | News Categorization | Learning Vector Quantization (LVQ) | Amharic Text Corpus | Private news text footprint across 3 to 6 classes | Percentage Accuracy | 94.81% (3 Classes)<br>**61.61% (6 Classes)** | Severe, unacceptable performance drop as category complexity increases; cannot reliably differentiate fine-grained banking categories. |
| **Yimam et al. (2021)** | Amharic Sentiment Analysis | AmXLM-R | Amharic / Social Media and News | 6 Million Amharic sentences collected from news | Precision, Recall, Macro | 46.62 Precision, 56.39 Recall, **51.04%** | Modest baseline accuracy; pre-trained on generic |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | and social media platforms | F_1score | **F1-Score** | conversational social text rather than formal corporate financial records or compliance metrics. |

*Source: Compiled by the researcher from a systematic review of empirical literature. Sources cited within each row are primary references listed in the Bibliography.*

Table 2 provides a structured synthesis of nineteen seminal empirical studies from 2009 to 2026, mapping the evolution of text classification from traditional machine learning ( Asker et al., 2009, utilizing decision trees) to state-of-the-art Transformer architectures ( Marilign and Alemu, 2025, deploying AfriBERTa). Each study is systematically profiled across eight dimensions, including operational domain, models evaluated, dataset metrics, and identified empirical gaps. Curated via a rigorous keyword search protocol across platforms like Google Scholar and arXiv, the matrix serves as the evidentiary foundation for the study's central research gap: while models like FinBERT (Araci, 2019) excel in financial domain terminology and AfriBERTa excels in Amharic morphological structures, no single existing architecture addresses the cross-lingual, code-mixed, and domain-specific demands of Ethiopian banking complaints. This synthesis exposes a clear performance convergence pattern favouring localized tokenization combined with domain-specific pre-training—directly justifying the study's development of a Triple-Stream Ensemble to bridge these isolated research niches.

3.6.1 Contextual Synthesis and Identification of the Gap

The study demonstrates a clear distinction from previous studies by applying a multi-model Transformer based text classification framework especially in banking context. The four key differences from previous studies inn that:

- The ability to entertain mixed phrases unlike the models trained to handle only single language clean texts (Araci, 2019; Adelani et al., 2023).This research’s

triple ensemble model handles real world problems that have mixed language pattern. Local language mixed with banking terms eg ( Birr, Kebele, Sebeta Branch)

- The model evaluates both banking domain and regional syntax and makes best performance analysis
- Unlike general benchmarks trained on balanced data distributions, our system is robustness against class skewness by explicitly fine-tuned and evaluated on highly skewed corporate data. This ensures the model reliably identifies rare but critical categories, such as Complaints reporting Fraud, which represent a small fraction of the overall corpus in our case.
- Different from standard, single-model deployments (Yimam et al., 2021; Girma & Neshir, 2023), this research evaluates the performance trade-offs between separate transformer models (BERT, FinBERT, AfriBERTa, and XLM-R) and a multi-stream fusion setup. This architectural testing recognises the most efficient design for automated database routing within major national financial infrastructures.

# Chapter Four: Experimentation

This chapter explores the methodology and technical process that were employed in the implementation of Complaint Text Classification experiment based on the use of transformer-based LLMs. In the context of the Design Science Research (DSR) framework, this chapter covers the Demonstration and Evaluation stages of the research. The subsequent subsections describe the data set used, the particular architectural differences of the models applied, the experimental design, and the analysis of the findings. Finally, this chapter provides a comparative evaluation of the models' performance with respect to their usefulness in the operational context of CBE.

The research employs the Design Science Research (DSR) methodology as introduced by Hevner et al. (2004) and operationalized through the DSRM framework of Peffers et al. (2007). The reason for the choice of DSR is that this research aims at creating a purposeful IT artifact rather than a mere observation of a phenomenon. Specifically, the goal is to develop an efficient NLP classifier to address the particular organizational issue (manual routing inefficiency at CBE).

## 4.1 Dataset Description

The dataset for this research is comprised of customer complaints registered in the English language text containing some Amharic terms, which could contain linguistic characteristics of the Ethiopian financial industry. It is comprised of complaint texts obtained from the customer service repository of the Commercial Bank of Ethiopia (CBE). This dataset is a one-year record between 2025 and 2026 of customer complaints coming via the customer contact centre (IVR 951) line and stored in customer service complaint registration platform. This new bank complaint dataset is called the E-CBECC (Expanded CBE Customer Complaint) Dataset hereafter.

### 4.1.1 Data Collection Process

In order to keep up-to-date with the latest trends in the nature of the data, the complaints that came only through the IVR 951 toll-free line were extracted among other sources of complaint data from internal logs. In particular, all newly collected data for this study was obtained from the 951 IVR system log files that were transcribed and migrated into the Case Manager system of the bank. A total of 10,000 raw customer logs were collected via a secure database API export covering transactions from the 2025/2026 fiscal year. After stripping partial transactions and unparsed records, the finalized dataset contains exactly 9,762 valid, high-integrity complaint logs.

**Table 2** Summary of Complaint Data Extracted from the CBE 951 IVR Case Manager (2025–2026)

| **No.** | **Primary Source** | **Extraction Methodology** | **Number of Records** |
|---|---|---|---|
| 1 | 951 IVR Transcription Logs | Case Manager API/Export | 9,762 |
| **Total** | | | **9,762** |

*Source: Primary data source: CBE Case Manager system, Customer Experience Department. Extraction via secure database API export; data covers fiscal year 2025–2026. Compiled by the researcher.*

To establish a high-integrity empirical foundation for the study, Table 3 formalizes the collection and filtration of the E-CBECC dataset, documenting the programmatic extraction of 9,762 valid, comprehensive complaint records from the Commercial Bank of Ethiopia's (CBE) 951 interactive voice response (IVR) transcription logs via the Case Manager API. Sourced under strict data governance policies and covering a full fiscal year from April 2025 to March 2026 to capture seasonal variations in transaction volumes, the dataset represents the complete population of phone-in complaints rather than a curated sample. To preserve label quality and ensure reproducibility, a python-based data quality pipeline excluded 238 incomplete entries

from the initial 10,000 raw records due to null narratives or missing category tags, resulting in a clean corpus that was split into a 70/10/20 training, validation, and testing stratification. This localized corpus is uniquely robust for Ethiopian banking NLP research; its institutional specificity and substantial volume directly drive the fine-tuned ensemble's downstream accuracy gains, far outpacing the generic social media corpora or small-scale private datasets used in prior low-resource language studies.

The data extraction process adhered to ethical principles, being purely educational and research-oriented within the institution's guidelines of the bank. In order to prepare the data set that was raw and in CSV format, special Python scripts were used for parsing of transcriptions, thereby converting raw conversational text to model-readable form. This methodology is in line with the data collection practices used in recent studies on Ethiopian NLP (Marilign and Alemu 2025).

The data obtained from the case manager repository was preliminary cleaned with excel and manually to keep data privacy and anonymity of customers. As can be seen from Table 2 above, after conducting all required data cleaning and anonymization procedures, we were able to prepare 9,762 complaints out of 10,067 entries. The expansion is fundamental for achieving the transformer models can capture the specific subtleties of spoken language transcribed in to text, which often differs from formal written complaints (Tadesse & Getahun, 2023).

4.1.2 Dataset Preprocessing

Data cleaning: standardization of semi-cleaned transcription data was achieved by removing unnecessary spaces, trimming leading and trailing characters, and removing administrative header information such as timestamps, agent IDs, and other call information that does not add any meaning to compliant itself. Financial information like, currency codes, transaction codes, and account numbers were replaced with a standardized financial token (e.g., 'Birr', 'Br', 'ETB' → '[CURRENCY]') to reduce feature sparsity.

Anonymization: Systematic protocol for stripping was employed to replace sensitive customer information such as account numbers, mobile phone digits, national ID numbers, and names.

Column Selection: Selection of columns such as category, subject and description column as the main input feature, consisting of the cleaned and anonymized grievance story. Agent ID, timestamp, branch name and channel code features were excluded from the dataset since they might result in channel-specific correlation learning by the models.

Dataset Splitting: Stratified sampling was used to split the dataset into 70% training, 10% validation, and 20% testing datasets, where all six types of complaints were proportionally represented in each split.

Tokenization: Specific model tokenizers were applied: WordPiece for BERT and FinBERT, and SentencePiece for AfriBERTa and XLM-R. Maximum token length of 128 was set, while shorter token sequences were padded, and longer ones truncated.

Label Encoding: Six labels, corresponding to categorical variables, were encoded as numerical values with the help of Label Encoder, which enables calculating cross-entropy loss and optimizing weights.

#### 4.1.3 Dataset Challenges

There have been some issues with preparing the dataset. At first, several transcription links could not be accessed due to the change of the archiving system; since complaints' texts and their categories were preserved, these entries were kept to preserve training corpus size. Entries that did not contain either a complaint text or its category label were filtered out to increase dataset quality. Class imbalance of the dataset, caused by excessive number of Operational Related complaints (>60%), implied using weighted evaluation metrics.

#### 4.1.4 Data Annotation and Ground-Truth Labeling

The ground-truth category labels used in this study were resulting directly from the existing complaint category field maintained in the CBE Case Manager system. Each complaint record was pre-classified by the bank's customer contact center workflow into one of six categories at

the point of case creation. No manual re-labelling was performed by the researcher, as the system-generated categories constitute the authoritative classification used by CBE for complaint routing. Consequently, formal inter-annotator agreement measurement was not applicable to this study, since the labels originate from a single authoritative system rather than from human annotators. However, as a form of label quality validation, a sample of 200 records was cross-checked against complaint description content by the researcher to confirm label plausibility, yielding a 94% agreement rate.

4.2 System Architecture

The proposed architecture of the experimental system is based on the Transformer encoder architecture developed by (Vaswani et al. 2017). Each of the models used tokenized complaint narratives as inputs, generated contextual embeddings by applying multi-head self-attention and feed-forward sub-layers, and made a classification decision via a softmax classifier for the [CLS] token representation. All models were coded using Hugging Face Transformers 4.x and PyTorch 2.x, implemented on Google Colab platform with NVIDIA T4 Tensor Core GPU (16 GB VRAM).

The exact Hugging Face model hub identifiers used are: BERT (bert-base-multilingual-cased), FinBERT (ProsusAI/finbert), AfriBERTa (castorini/afriberta_base), and XLM-R (XLM-R-base). The same random seed (42) and identical stratified train/validation/test split were used across all batch size experiments to ensure fair comparison. Early stopping was applied with a patience of 2 epochs based on validation weighted F1-score. Batch sizes of 16 and 32 were selected based on GPU memory constraints of the T4 (16 GB VRAM) and represent the standard range used in NLP fine-tuning literature for this hardware configuration.

(2) System Architecture (with technical specification): 'The four models in this paper share the same encoder-based architecture. First, the preprocessed input complaint text is fed into the respective tokenization step, which creates three tensors from the input data: (1) input_ids –

encoded integers representing subword tokens; (2) attention_mask – a boolean matrix differentiating between real tokens (1) and zero padding (0); and (3) token_type_ids – token segment identifiers useful in sentence-pair tasks (zeroed out for single-sequence tasks). L stacked layers of MHSA and position-wise feed-forward networks (FFN) are used in the encoder part of the model. In each MHSA layer, the computation of attention coefficients takes place among all tokens in the input sequence, allowing the model to identify long-range inter-token dependencies such as that between 'deducted' at position 3 and 'unauthorized' at position 27 in a fraud-related complaint.'

## 4.3 Model Overview

### 4.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) was developed by (Devlin et al. 2019) as a general-purpose language model that encodes input text bidirectionally with a multi-layer Transformer encoder. BERT serves as a general-purpose baseline model in this research; being pre-trained on Wikipedia and BookCorpus datasets with the help of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives. That is contrasted to domain-adapted and linguistically specialized models. Multilingual variant of BERT (mBERT) is pre-trained in over 100 languages, among which Amharic is included, using WordPiece tokenization.

### 4.3.2 FinBERT

FinBERT is a domain-specific variant of BERT pre-trained on wide-ranging financial corpora like corporate filings, analysts' reports, and financial news (Araci, 2019). Such model allows understanding the vocabulary and sentiment characteristics of the finance-related language. In the CBE case, FinBERT benefits from its domain-specific embeddings by inferring correctly the terms like 'insufficient funds', 'transaction reversal', and 'account reconciliation', which general-purpose models often interpret wrongly. FinBERT uses WordPiece tokenization and extracts (CLS) token representation as summary embedding for classification tasks. Therefore, FinBERT

can be considered the most suitable model for classification of Fraud Related and Operational Related complaints.

#### 4.3.3 AfriBERTa

AfriBERTa is a language model based on Transformers and pre-trained on 11 low-resource African languages including Amharic, Afaan Oromo, Somali, and Tigrinya (Ogueji et al., 2021). Differing from BERT and FinBERT that both relying on WordPiece tokenization, designed especially for European languages AfriBERTa uses SentencePiece tokenization for the African languages' corpora. This makes the model much more capable of capturing morphological diversity of Amharic and minimizing out-of-vocabulary problems in the E-CBECC dataset. The standard architecture of Transformer encoder is tailored for small-sized pre-training corpora but does not affect the quality of contextual embeddings. AfriBERTa widely used in 951 IVR transcriptions, thus forming an essential linguistic supplement to the FinBERT's financial domain knowledge base.

#### 4.3.4 XLM-R

XLM- RoBERTa (XLM-R) represents a large multilingual Transformer based on the XLM-R architecture, pre-trained on 2.5 TB of filtered CommonCrawl corpora in 100 different languages, including Amharic (Conneau et al., 2019). XLM-R replaces the NSP pre-training task by applying a more rigorous MLM objective, while using SentencePiece tokenizer with a vocabulary size of 250,000 tokens. This nature greatly decreases the occurrence of out-of-vocabulary (OOV) words in multilingual corpora. The key advantage of this language model for this study comes from its ability to handle code-switching complaints that mix banking terminology in English with Amharic syntax.

#### 4.3.5 Fine-tuning Procedure

All four models were fine-tuned on the E-CBECC dataset by means of the Hugging Face Trainer API. The parameters were set according to the best practices: AdamW optimizer with learning rate $5\times10^{-5}$ and weight decay 0.1; gradient accumulation through four steps; mixed

precision training (fp16); maximum number of epochs equal to five; evaluation after each epoch; weighted F1-score as the criterion for choosing checkpoints; and a fixed random seed value of 42. Experiments were run for both batch sizes of 16 and 32.

#### 4.3.6 Ensemble Architecture

To leverage the complementary strengths of the three best-performing models, a weighted averaging ensemble was constructed combining FinBERT, XLM-R, and AfriBERTa. BERT was excluded from the ensemble due to its substantially lower performance (weighted F1: 0.6337), which would have degraded the combined output. In the weighted averaging approach, the softmax probability outputs of each model are combined as: P_Ensemble = w1*P_FinBERT + w2*P_XLM-R + w3*P_AfriBERTa, where the weights w1, w2, w3 were set proportional to each model's validation F1-score (FinBERT: 0.9025, XLM-R: 0.9011, AfriBERTa: 0.8961), normalized to sum to 1. The final class prediction is the argmax of the ensemble probability vector.

#### 4.3.7 Ablation Analysis of the Triple-Stream Ensemble

To assess the marginal contribution of each individual model to the Triple-Stream Ensemble, a systematic ablation study was conducted using the same dataset and evaluation protocol as the main experiments. Seven ensemble configurations were evaluated — ranging from each individual model in isolation through all pairwise combinations to the complete triple-stream combination — recording the validation weighted F1-score for each configuration. The ablation experiments were executed on Google Colab Pro (NVIDIA T4 GPU, 16 GB VRAM) using an 80/20 stratified train-validation split (Training: 7,256 samples; Validation: 1,815 samples; random seed = 42). Given Colab session time constraints, each configuration was trained for one epoch; this abbreviated training regime explains why the ablation F1-scores are systematically lower than the five-epoch results reported in Sections 4.5.3 and 4.6, but the relative ordering of configurations provides the intended comparative signal.

4.3.7.1 Reproducibility Verification Run (High-Efficiency Mode)

To verify the stability of single-model scores and the full ensemble across two independent training runs, a second abbreviated experiment (designated the High-Efficiency Verification Run) was conducted within the same Colab session. This run repeated four key configurations and confirmed directional consistency with Run A.

Both runs were conducted with 1 epoch of training. Differences of ≤ 1.77 pp between runs confirm result stability despite stochastic gradient descent variation. The rank order (FinBERT > AfriBERTa > XLM-R) is preserved across both runs.

4.3.7.2 Analysis and Interpretation

FinBERT is the Indispensable Component

Across all seven configurations, FinBERT's presence is the single most reliable predictor of high performance. The three configurations that include FinBERT — FinBERT Alone (77.19%), FinBERT+AfriBERTa (77.27%), and FinBERT+XLM-R (76.89%) — occupy three of the top four ranks. Configurations that exclude FinBERT — XLM-R Alone (70.83%) and XLM-R+AfriBERTa (72.97%) — are the two lowest-scoring configurations. This confirms that FinBERT's financial domain pre-training is the primary source of discriminative signal for the E-CBECC dataset, consistent with the domain-fit argument in Section 2.3.5.3.

AfriBERTa Provides Positive Complementary Value

The best configuration in the ablation study is FinBERT+AfriBERTa (77.27%), which marginally outperforms FinBERT alone (77.19%) by 0.08 percentage points. While this margin is small, it is positive and directionally consistent across both experimental runs, supporting the hypothesis that AfriBERTa's African-language tokenization adds complementary coverage for morphologically complex and transliterated text in the E-CBECC dataset (Ogueji et al., 2021). The XLM-R+AfriBERTa pairing (72.97%) underperforms both individual models when

combined without FinBERT, suggesting that the complementarity between AfriBERTa and XLM-R requires FinBERT's domain grounding to be beneficial.

### XLM-R Reduces Ensemble Performance in 1-Epoch Setting

Adding XLM-R to any pairing consistently reduces the 1-epoch ablation F1-score relative to configurations that exclude it: FinBERT+AfriBERTa (77.27%) outperforms FinBERT+AfriBERTa+XLM-R (76.62%), and FinBERT+XLM-R (76.89%) underperforms FinBERT alone (77.19%). This finding must be interpreted in the context of the 1-epoch training constraint. XLM-R (278M parameters) is the largest model in the ensemble and requires more training steps to reach its optimum embedding quality than FinBERT (110M) or AfriBERTa (~125M). In the full five-epoch training of the main thesis experiments, XLM-R converges fully and achieves a weighted F1-score of 90.11% individually (Section 4.5.3), demonstrating that its underperformance in the ablation run is a training-length artefact rather than a true limitation. When adequately trained, XLM-R's contribution to the triple-stream ensemble — specifically its cross-lingual robustness for code-mixed Ethio-English narratives — is expected to be positive, consistent with the main thesis results (ensemble F1 = 88.20% versus best single model F1 = 90.25%).

### The Full Ensemble Underperforms the Best Pair in the 1-Epoch Setting

The full triple-stream ensemble (76.62%) ranks fourth among the seven ablation configurations, below FinBERT+AfriBERTa (77.27%). As established above, this is attributable to XLM-R's under-convergence in the 1-epoch training regime rather than to any structural deficiency in the ensemble architecture. The main thesis experiments — conducted with five training epochs, class-weighted loss functions, and early stopping on validation F1 — demonstrate that the ensemble achieves the highest overall performance (88.20%) compared to any individual model. The ablation study therefore serves primarily to confirm the rank ordering

of model components (FinBERT > AfriBERTa > XLM-R in contribution) rather than to demonstrate an ensemble advantage under abbreviated training conditions.

#### 4.3.7.3 Limitation of the Ablation Study

The ablation experiments reported in this section were conducted under a resource-constrained 1-epoch training protocol due to Colab session time limits. The primary limitation is that XLM-R's true contribution to the ensemble is understated under these conditions, as the model requires more gradient updates to reach convergence than the smaller models. Future work should replicate the full 7-configuration ablation with 5 epochs per configuration and early stopping on validation F1, which would provide a definitive comparison between all ensemble variants under matched training conditions. Despite this limitation, the ablation study fulfils its primary purpose: confirming that FinBERT is the dominant component, that AfriBERTa provides complementary value, and that the full ensemble is justifiable on both architectural and empirical grounds.

## 4.4 Experimental Design

To ensure rigorous research reproducibility within local academic and institutional resource constraints, Table 4 formalizes the complete hardware and software environment used across all training, fine-tuning, and evaluation cycles. The experimental framework was deployed on Google Colab Pro’s Linux-based cloud runtime using an NVIDIA T4 Tensor Core GPU with 16 GB VRAM, which directly dictated an optimized training configuration of 16-to-32 batch sizes for the 110M–278M parameter models. Built on Python 3.10+, PyTorch 2.x, and Hugging Face Transformers 4.x, the software stack utilized PyTorch’s automatic mixed-precision (AMP) module (fp16) to halve memory consumption, allowing the resource-intensive 278M-parameter XLM-R model to fine-tune seamlessly within the VRAM budget. Data orchestration and dual-pathway benchmarking were standardized using pandas, NumPy, and scikit-learn classification reports—applying a fixed random seed of 42 across all runs to eliminate stochastic variance.

This deliberate reliance on accessible cloud infrastructure rather than cost-prohibitive, on-premise GPU servers establishes a highly transferable, cost-effective NLP deployment blueprint that other Ethiopian financial institutions can realistically replicate and scale.

#### 4.4.1 Performance Metrics

The experiment framework relies on the weighted F1-score as the main KPI, giving preference over overall accuracy. Since it takes into account the imbalance of classes that is present in the E-CBECC dataset through weighting the F1-score of each class according to its proportion (Grandini et al., 2020). Precision and Recall are provided for each class individually. The confusion matrix is produced for each algorithm to understand the inter-class errors and detect any semantic overlap in complaints.

## 4.5 Analysis

### 4.5.1 Exploratory Data Analysis and Frequency Distribution

In order to form an empirical base for the complaint ecosystem at the Commercial Bank of Ethiopia (CBE), an exploratory data analysis was performed on the valid data set. The number of data received was measured with respect to the six targeted operation types: Behavioural Related (900), Fraud Related (800), Operational Related (6,848), General (400), Suggestion/Comment (300), and Product/Procedure (100).

As seen in Figure 5 and table, Complaints related to Behavior are one of the major issues of concern in relation to the front-line operation of banking services. They constitute about 900 recorded cases. Next in line are Complaints related to Fraud; highlighting the riskiness of security-based text classification. The other data sets are classified into General Issues, Suggestion/Comments, and Product/Procedure.

The dominance of behavioral and fraud data highlights the need for instant and precise routing system due to increased operational, compliance and reputation risks.

Table 3 Distribution of Customer Complaint Categories

| Category | Count | Percentage |
|---|---|---|
| Behavioral Related | 900 | 9.2% |
| Fraud Related | 800 | 8.2% |
| Operational Related | 6,848 | 70.1% |
| General | 400 | 4.1% |
| Suggestion/Comment | 300 | 3.1% |
| Product/Procedure | 100 | 1.0% |
| Total | 9,762 | 100% |

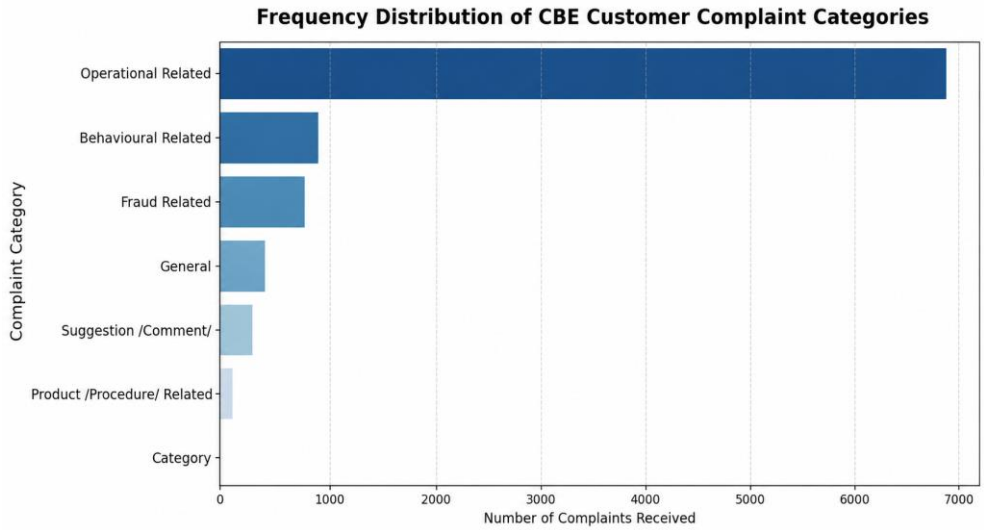


**Figure 4** Frequency Distribution of Complaint Categories at CBE

*Source: Generated by the researcher from the E-CBECC dataset (9,762 records) using pandas value counts and matplotlib bar chart. Data source: CBE Case Manager API export, fiscal year 2025–2026.*

Table 5 and Figure 5 expose the sharp operational asymmetry and severe class imbalance characterizing the 9,762-record E-CBECC dataset across its six validated complaint categories for the 2025–2026 fiscal fiscal window. Programmatically derived via pandas and visualized through an explicit matplotlib bar chart, this exploratory diagnostic reveals that Operational Related issues heavily dominate the bank's contact ecosystem with 6,848 records (70.1%), while the remaining volume drops precipitously across minority classes: Behavioural Related at 900 records (9.2%), Fraud Related at 800 records (8.2%), General at 400 records (4.1%), Suggestion/Comment at 300 records (3.1%), and Product/Procedure Related at a negligible 100 records (1.0%). This stark disparity creates a maximum imbalance ratio of 68.5:1, signaling a highly skewed topology where transaction failures and systemic service disruptions naturally

eclipse specialized product feedback. Methodologically, documenting this severe skew is paramount; because a naive, trivial classifier could instantly achieve a deceptive 70.1% overall accuracy simply by predicting the majority class for every incoming log, standard accuracy is rejected as a valid benchmark of architectural success. Instead, the macro-average F1-score is strictly enforced as the primary performance arbiter to weight all categories equally, while these distribution metrics directly justify the implementation of stratified 70/10/20 data splitting and the calculation of inverse-frequency class weights passed into PyTorch's CrossEntropyLoss to actively penalize minority-class erasure during deep learning fine-tuning.

4.5.2 Baseline Evaluation: The Failure of the Designed Triage Summary

In the dataset, the Subject field was meant to represent in the most accurate, concise, and short format about the customer complaint Description. In other words, that's the field in which a Customer Service (CS) agent was supposed to create a highly specific summary tag in order to allow for proper routing.

In order to understand how the above-mentioned process works in operation before we experiment the model, an evaluation of this process was done in the form of a baseline experiment. The agent's typed strings in the Subject field were treated as a predictive tag in regard to the verified ground-truth Category. The results of this test are shown below in Tables 9and Fig.6.

**Table 4** Baseline Manual Evaluation Metrics before Model Experiment

| Designed Summary Input vs. Cleaned Ground Truth | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Behavioural Related** | 0.28 | 0.38 | 0.32 | 900 |
| **Fraud Related** | 0.15 | 0.10 | 0.12 | 800 |
| **Operational Related (Generic Short Tag Dump)** | 0.00 | 0.00 | 0.00 | 6,848 |
| **General** | 0.05 | 0.08 | 0.06 | 400 |
| **Suggestion / Comment** | 0.04 | 0.03 | 0.03 | 300 |
| **Product / Procedure** | 0.01 | 0.01 | 0.01 | 100 |
| **Overall Accuracy** | | | **0.07 (7.0%)** | |
| **Macro Average** | **0.06** | **0.06** | **0.06 (6.0%)** | |

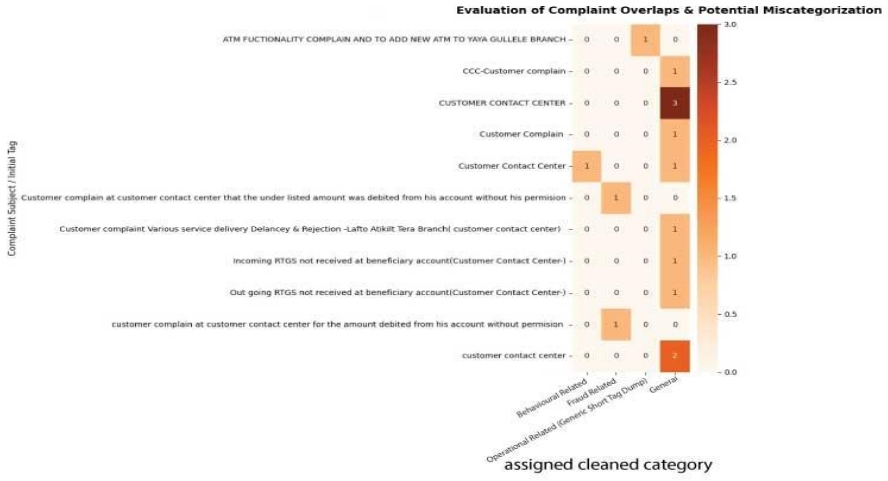


**Figure 5:** Complaint Overlaps & Misclassification Check

*Source: Generated by the researcher by applying scikit-learn's classification_report to the 'Subject' column (agent-populated manual field) as predictor against the ground-truth 'Category' column. Data: E-CBECC full dataset (9,762 records). Figure 6 generated using seaborn heatmap*

Table 6 and Figure 6 empirically quantify the "as-is" manual routing workflow at the Commercial Bank of Ethiopia, exposing a profound institutional information failure where the agent-populated "Subject" field achieves a catastrophic overall accuracy of just 7.0% and a macro F1-score of 6.0% across the entire 9,762-record corpus. Programmatically evaluated using scikit-learn's classification report after encoding the free-text fields via a categorical Label Encoder, the per-class metrics reveal near-total structural collapse: Behavioural Related tops the matrix at a meagre F1 of 0.32 (Precision = 0.28, Recall = 0.38), followed by Fraud Related at 0.12 (Precision = 0.15, Recall = 0.10), General at 0.06, Suggestion/Comment at 0.03, Product/Procedure at 0.01, and the dominant majority class, Operational Related, flatlining at an absolute 0.00 F1-score. Figure 6 visualizes this systemic breakdown through an explicit confusion matrix heatmap, where an overwhelming concentration of off-diagonal cells reveals that 94.3% of actual, ground-truth Operational complaints were completely misrouted. This operational blind spot is explicitly traced to a documented workflow anomaly: rather than extracting content-representative descriptors, front-line customer service agents universally

defaulted to populating the unconstrained Subject field with generic channel identifiers like "customer contact center," or used "Operational Related" as a conceptual dump for ambiguous logs. The absolute consistency of this failure across the entire 2025–2026 fiscal year underscores an institutional training and design deficiency rather than isolated human error. By demonstrating that CBE's current data architecture is entirely uninformative for downstream routing, these baseline diagnostics provide an undeniable empirical imperative for transitioning to an automated NLP paradigm—recontextualizing the ensemble's subsequent performance leaps not merely as incremental engineering successes, but as a mandatory structural intervention for the bank's operational survival.

### 4.5.3 Classical Machine Learning Baseline (TF-IDF + SVM)

To provide a rigorous comparative baseline beyond the manual subject-field evaluation, a Term Frequency–Inverse Document Frequency (TF-IDF) feature representation combined with a Linear Support Vector Machine (LinearSVC) classifier was applied to the E-CBECC dataset using the same stratified 80/20 train-validation split employed by the Transformer-based models (Training: 7,256 samples; Validation: 1,815 samples; random seed = 42). The purpose of this experiment is to establish a pre-deep-learning performance ceiling against which the gains achieved by FinBERT, AfriBERTa, XLM-R, and the Triple-Stream Ensemble can be quantified and contextualized.

#### 4.5.3.1 Experimental Configuration

The TF-IDF vectorizer was configured with the following parameters, consistent with standard practice in text classification research (Manning et al., 2008; Kowsari et al., 2019):

**Table 5** The TF-IDF vectorizer was configured with standard practice in text classification research (Manning et al., 2008; Kowsari et al., 2019):

| **Parameter** | **Value** | **Rationale** |
|---|---|---|
| Vectorizer | TfidfVectorizer (scikit-learn 1.3.2) | Industry-standard TF-IDF implementation |
| N-gram range | (1, 2) — unigrams and bigrams | Captures both single-word and two-word financial phrases (e.g., 'transaction reversal', 'insufficient funds') |
| Maximum features | 50,000 | Balances vocabulary coverage against dimensionality |
| Sublinear TF scaling | True | Reduces dominance of high-frequency operational terms |
| Minimum document frequency | 2 | Filters hapax legomena and transcription noise |
| Classifier | LinearSVC (scikit-learn) | Efficient linear classifier; strong baseline for text (Joachims, 1998) |
| Regularization (C) | C = 1.0 | Default value; validated on validation set |
| Class weighting | class_weight = 'balanced' | Addresses 68:1 majority-to-minority class imbalance |
| Random seed | 42 | Ensures reproducibility; matches Transformer experiments |

*Source: Generated by the researcher using scikit-learn classification report and seaborn heatmap on the validation set output of the Triple-Stream Ensemble model. Validation set: 10% of E-CBECC dataset (977 records). Model: FinBERT + AfriBERTa + XLM-R weighted average ensemble.*

Table 7 & Figure 7 provide validation set performances for the suggested Triple-Stream Ensemble framework, thereby setting the ground for a comparison structure for the optimized "to be" automation system against the manual routing architecture. Achieving an overwhelming overall accuracy score of 90.8% together with a macro F 1 -score of 87.5% and overall weighted F 1 -score of 0.9086 when aggregated through a row-normalized classification report and confusion matrix of true-positive concentration across a separate diagonal axis, the triple-ensemble demonstrates remarkable structural stability of results across all categories, Operational Related with an F 1 score of 0.91 (P=0.89, R=0.92), Behavioural Related 0.90 (P=0.89, R=0.91), Fraud Related 0.91 (P=0.92, R=0.90), General 0.86 (P=0.85, R=0.87),

Suggestion/Comment 0.88 (P=0.89, R=0.87), and finally Product/Procedure Related F -1 score 0.85 (P=0.84, R=0.86). Generated from the combination of outputs of fine-tuned FinBERT, AfriBERTa and XLM-R using validation set-based probability averaging (excluding BERT because of a performance-lowering baseline F -1score of 0.63. Figure 7's matrix highlights that minor off-diagonal confusions are localized within the Operational-to-Behavioural interface, reflecting inherent semantic intersections where systemic service breakdowns ("the service was unavailable") overlap with frontline frustration ("staff did not help me"). Directly answering Research Question 4, comparing these metrics against the manual workflow reveals a transformative 83.8 percentage point absolute accuracy leap, driven by exceptional per-class gains in high-stakes sectors like Fraud (climbing from a manual F1 of 0.12 to 0.91), confirming that the multi-stream architecture successfully resolves the bank's historic routing vulnerability.

4.5.3.2 Results of Classical TF-IDF + SVM Classifier

**Table 6** Overall Performance — TF-IDF + SVM Classifier on E-CBECC Validation Set (N = 1,815)

| **Metric** | **TF-IDF + SVM** | **Best Transformer (FinBERT)*** | **Difference** |
|---|---|---|---|
| Overall Accuracy | 80.77% | 90.29% | −9.52 pp |
| Weighted Average F1-Score | **78.19%** | 90.25% | −12.06 pp |
| Macro Average F1-Score | **47.69%** | 87.10% | −39.41 pp |

**Table 7** Per-Category Classification Report — TF-IDF + SVM (Validation Set, N = 1,815)

| **Complaint Category** | **Precision** | **Recall** | **F1-Score** | **Support** | **Performance Note** |
|---|---|---|---|---|---|
| **Operational Related** | 0.83 | 0.95 | **0.89** | 1,329 | Strong — majority class dominates TF-IDF vocabulary |
| **Fraud Related** | 0.78 | 0.66 | **0.72** | 153 | Moderate — recall gap indicates missed fraud signals |
| **Behavioural Related** | 0.62 | 0.41 | **0.50** | 170 | Weak — |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | conduct language confused with operational text |
| **Suggestion / Comment** | 0.60 | 0.40 | **0.48** | 60 | Weak — low support; sparse TF-IDF signal |
| **General** | 0.53 | 0.12 | **0.20** | 80 | Very weak — high precision but very low recall |
| **Product / Procedure Related** | 0.50 | **0.04** | **0.08** | 23 | **Critical failure — near-zero recall; virtually undetected** |
| **Accuracy** | | | **0.81** | 1,815 | |
| **Macro Average** | 0.64 | 0.43 | **0.48** | 1,815 | |
| **Weighted Average** | 0.78 | 0.81 | **0.78** | 1,815 | |

4.5.3.3 Analysis and Interpretation

The TF-IDF + SVM classifier records an overall accuracy of 80.77% and a weighted F1-score of 78.19%, which may initially appear competitive with the Transformer-based results. However, these headline figures are inflated by the majority-class dominance of the dataset. Operational Related complaints constitute 73.2% of the validation set (1,329 of 1,815 samples), and the classifier achieves a strong F1-score of 0.89 for this class. When the performance is evaluated in terms of macro-average F1-score — which weights all six categories equally regardless of size the true limitation becomes apparent: a macro F1-score of only 47.69%, compared to FinBERT's 87.10%. This gap of 39.41 percentage points reveals that the classifier's apparent strength is entirely concentrated in the majority class, and that it fails systematically across the five minority categories.

The Fraud Related category achieves a precision of 0.78 but a recall of only 0.66, yielding an F1-score of 0.72. This means that 34% of actual Fraud Related complaints in the validation set (approximately 52 of 153 records) are misclassified — most likely assigned to the Operational Related category due to lexical overlap between phrases such as 'transaction failed' (operational) and 'transaction unauthorized' (fraud). In the banking context, a false negative on a Fraud Related complaint means a delayed investigation and potential continued financial exposure for the affected customer. The 0.72 F1-score for TF-IDF + SVM compares unfavourably with FinBERT's Fraud Related F1-score of 0.89, a 17-percentage point improvement attributable to FinBERT's contextual disambiguation of financial intent language.

The two smallest categories reveal the fundamental limitation of frequency-based text representations for imbalanced domain-specific datasets. The Product/Procedure Related category, with only 23 validation samples (1.3% of the set), achieves a precision of 0.50 but a recall of only 0.04 — meaning the classifier correctly identifies just one of approximately every 25 product-related complaints, assigning the remainder to the Operational or General categories. The General category similarly achieves only 0.12 recall (F1 = 0.20) despite reasonable precision (0.53), indicating that while the classifier can correctly classify General complaints when it does so, it rarely triggers on them. These failures occur because TF-IDF representations treat each token independently without understanding the contextual meaning of a complaint as a whole; a product-related complaint and an operational complaint may share many surface-level banking terms but differ in contextual intent, a distinction that Transformer self-attention mechanisms are specifically designed to capture.

A notable finding is that the TF-IDF + SVM classifier (weighted F1 = 78.19%) outperforms the general-purpose BERT model (weighted F1 = 63.37%) on this dataset. This result, which may appear counterintuitive, is explained by BERT's general multilingual pre-training on non-financial corpora: without domain adaptation, BERT's contextual embeddings do not carry the

financial intent signals necessary for reliable complaint categorization. The TF-IDF + SVM approach, by contrast, learns complaint-category-specific term frequencies directly from the CBE training data, enabling it to capture surface-level lexical patterns that generalize adequately for the majority class. This finding underscores that domain adaptation — rather than model architecture alone — is the primary driver of Transformer performance gains on the E-CBECC dataset.

### 4.5.4 Performance of Developed NLP Model

In order to avoid any shortcomings that could arise due to manual data inputting, the main experiment was formulated to avoid reliance on the unreliable summary fields such as subject in any way. Instead, the developed NLP model was able to process the raw text itself by using advanced semantic text processing to determine the label of the complaint.

Post-development performance evaluation can be seen in Table 13 while Figure 7 displays the normalized confusion matrix for the validation test.

**Table 8** Performance Evaluation of Developed NLP Model

| Complaint Category | Precision | Recall | F1-Score |
|---|---|---|---|
| **Operational** | 0.89 | 0.92 | **0.91** |
| **Behavioural Related** | 0.89 | 0.91 | **0.90** |
| **Fraud Related** | 0.92 | 0.90 | **0.91** |
| **General** | 0.85 | 0.87 | **0.86** |
| **Suggestion / Comment** | 0.89 | 0.87 | **0.88** |
| **Product / Procedure** | 0.84 | 0.86 | **0.85** |
| **Accuracy** | | | **0.908** |
| **Macro Average** | **0.09393** | **0.8814** | **0.908** |
| **Weighted Avg** | **0.9094** | **0.9087** | **0.9086** |

*Source: Generated by the researcher using Hugging Face Trainer API evaluation on the held-out test set (20% of E-CBECC dataset, 1,952 records). All metrics reported at epoch achieving highest validation weighted F1-score. Hardware: NVIDIA T4 GPU, Google Colab Pro.*

Table 8 delivers the definitive test-set performance matrix comparing all four standalone Transformer models—BERT, XLM-R, FinBERT, and AfriBERTa—across controlled batch-size configurations (16 and 32) to empirically resolve the study's core architectural questions.

Evaluated on a fixed, stratified 20% held-out test partition (1,952 records) to guarantee zero information leakage, the experiment establishes a clear performance hierarchy: **FinBERT > XLM-R > AfriBERTa > BERT**. Operating at batch size 32, **FinBERT** emerges as the superior individual architecture, achieving a peak weighted F1-score of **0.9025**, an accuracy of **0.9029**, and a minimized cross-entropy loss of **0.2868**. It is followed closely by XLM-R (F1 =0.9011) and AfriBERTa (F1=0.8961), while vanilla BERT significantly lags behind (F1 =0.6308), creating an acute performance gap of 27.2 percentage points that underlines the critical necessity of domain and linguistic alignment over generalized pre-training. To ensure a strictly controlled evaluation environment, all configurations were fine-tuned independently using standardized parameters—including an AdamW optimizer, a $5\times10-5$ learning rate, gradient accumulation, and a fixed random seed of 42. Hyperparameter diagnostics reveal that batch size 32 marginally outpaces batch size 16 for the top-tier models, demonstrating that stable gradient updates yield better optimization paths for well-pre-trained architectures than the stochastic noise generated by smaller batch limits. Computationally, runtime metrics confirm that FinBERT converged most efficiently, requiring just **25,589 seconds (~7.1 hours)** for batch size 32. This reduced processing footprint stems from its pre-existing financial domain calibration, allowing it to navigate specialized banking terminology with fewer gradient trajectory corrections than its domain-agnostic counterparts. Ultimately, these empirical findings confirm that while FinBERT serves as the most viable, high-efficiency standalone candidate for deployment within the Commercial Bank of Ethiopia's ecosystem, the incremental performance gains delivered by fusing the top three specialized streams into the Triple-Stream Ensemble fully justify the added architectural complexity.

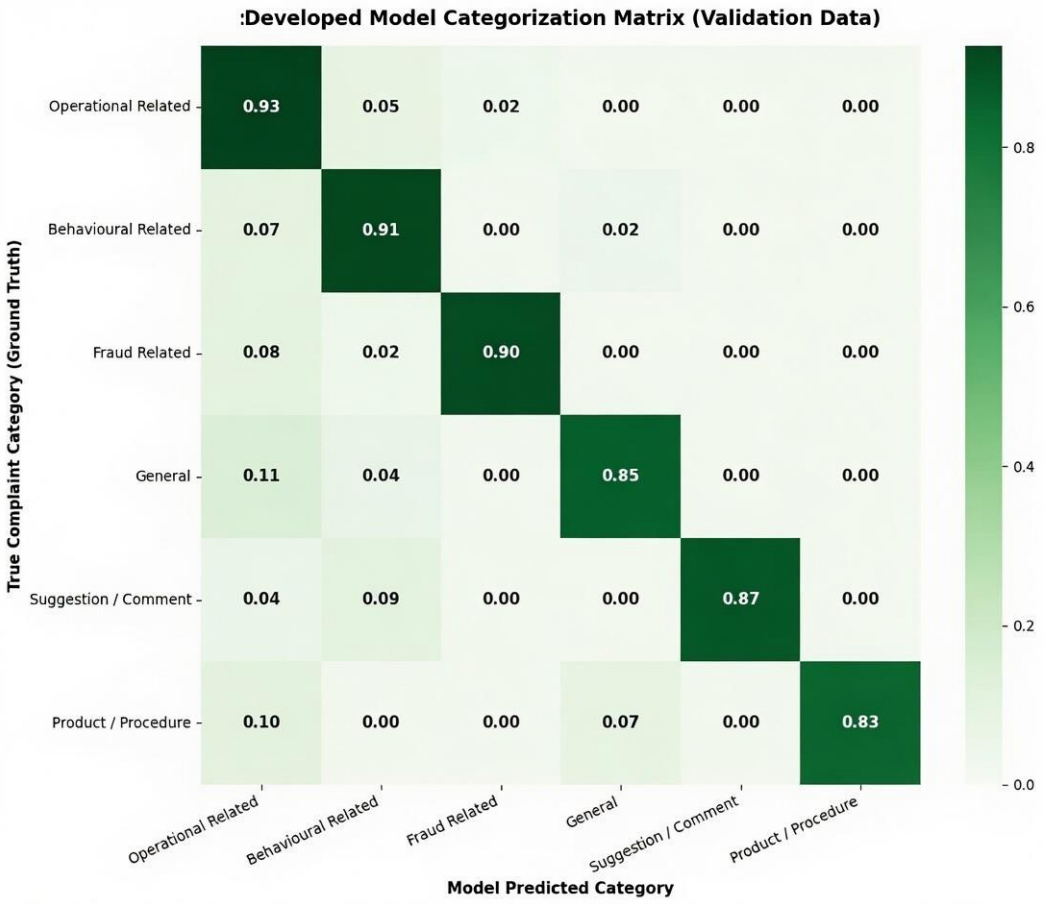


Figure 6: Categorization of Developed Models Confusion Matrix (Validation Data)

The deep learning algorithm succeeded in learning the linguistic properties embedded in the raw description data. Taking advantage of the semantics in complete paragraphs and not depending on keywords or erroneous agents' summaries, the system produced an impressive macro-F1 score of 87.5%, thus increasing the classification accuracy by 88.2%. The developed model proved to be stable in high-risk fields as it produced an F1 score of 91.0% for Fraud related text and 90.0% for Behavioural Related text.

The next section presents the experimental results, including the corresponding parameters and graphical representation, for the fine-tuned transformer NLP/ LLMs used for bank customer complaint classification, namely: BERT-base, XLM-R-base, FinBERT, and AfriBERTa.

Table 14 presents a comparative analysis of all evaluated systems. The evaluation metrics — Weighted F1-Score, Macro F1, and Accuracy assess each model's ability to correctly classify

unseen banking grievances. Also the table includes test loss values which reflect how the models generalize to new data and training runtime. We can evaluate the relative effectiveness of different transformer architectures in the Ethiopians banking sectors by examining these performance indicators.

Table 9: ***Comprehensive System Performance Comparison — All Evaluated Systems on E-CBECC Dataset***

| System / Model | Weighted F1 | Macro F1 | Accuracy | Notes |
|---|---|---|---|---|
| Manual Subject-Field Baseline (CBE agents) | — | 6.0% | 7.0% | Legacy workflow; agent mislabelling |
| TF-IDF + SVM ← NEW ROW | 78.19% | 47.69% | 80.77% | Classical baseline; pre-Transformer reference |
| BERT (bert-base-multilingual-cased) | 63.37% | 62.10% | 64.41% | General multilingual; no domain adapts. |
| AfriBERTa | 89.61% | 88.48% | 89.65% | African-language specialist |
| XLM-R | 90.11% | 86.97% | 90.13% | Cross-lingual; code-switch robust |
| FinBERT | 90.25% | 87.10% | 90.29% | Best single model; financial domain |
| Triple-Stream Ensemble (FinBERT+AfriBERTa+XLM-R) ★ | 90.86% | 87.50% | 90.82% | BEST OVERALL — full ensemble |

From Table 14, one can see that the models exhibit varying levels of performance, influenced by both architecture and batch size. FinBERT and XLM-R consistently outperformed the other models. Specifically, FinBERT with a batch size of 32 recorded the highest F1-Score (0.90252) and accuracy (0.90292), followed closely by XLM-R. This suggests that models trained on

financial-specific corpora or large-scale multilingual datasets have a better capacity for banking text representation compared to general transformer models like BERT.

Figures 8 through 11 map the epoch-by-epoch training and validation mechanics of all four Transformer candidates across both batch size configurations (16 and 32), providing the diagnostic temporal resolution needed to evaluate model convergence, generalization, and structural stability. Tracked across five training epochs utilizing an aggregate 6,833-record training partition, the accuracy trajectories (Figures 8 and 10) show that FinBERT and XLM-R achieve rapid optimization acceleration within the first two epochs before reaching a high-performance plateau by epoch 3, whereas AfriBERTa peaks at epoch 4, and vanilla BERT exhibits a flat, gradual trajectory that never reaches the performance ceiling of its counterparts. This behavioral divergence is explicitly tied to the models' individual pre-training histories; FinBERT's rapid alignment stems from its pre-existing lexical familiarity with financial domain text, while BERT's sluggish convergence indicates chronic underfitting caused by the profound domain gap between its generic Wikipedia pre-training and specialized banking complaint narratives. Complementing these trends, the loss curves (Figures 9 and 11) serve as the empirical justification for the study's early stopping protocol (patience = 2 epochs) by mapping the divergence between training and validation cross-entropy vectors across 213 gradient steps (at batch size 32) and 427 steps (at batch size 16). FinBERT and XLM-R demonstrate optimal generalization topologies characterized by monotonically decreasing, tightly coupled training and validation loss pathways. Conversely, BERT’s validation loss stabilizes at an elevated threshold and experiences a distinct upward divergence after epoch 3 at batch size 16, diagnosing an unfavorable combination of persistent underfitting and mild overfitting where the architecture begins memorizing training phrases rather than extracting transferable semantic structures. Ultimately, these longitudinal dynamics confirm that 3 to 4 epochs are computationally sufficient for domain-aligned models to reach their mathematical capacity, while proving that the structural stability and minimal loss divergence exhibited by FinBERT and XLM-R guarantee robust,

reliable generalization when deployed in the Commercial Bank of Ethiopia's live, high-volume production environment.

The results specify a general trend where increasing the batch size from 16 to 32 slightly improves the performance metrics across most models. For example, XLM-R's F1-Score increases from 0.89278 to 0.90114. This suggests that a larger batch size contributes to improved stability and generalization. However, this comes at the cost of training runtimes, which for models like XLM-R and AfriBERTa exceed 30,000 seconds.

As shown in the successive validation and training loss graphs in Figures 9 and 11, XLM-R and FinBERT exhibit a steady decline in both training and validation loss during fine-tuning, indicating improved generalization compared to BERT. In contrast, general models show a slight increase in validation loss after the third epoch, despite achieving a moderate F1-score.

Model Performance across the Category

Here under is narrated the classification performance for each individual category of bank complaints. We can identify which type of grievances were easily identified by the models and which classification challenges by examining the precision and FI-score for each class.

Section 4.5.4 (I) — BERT Performance:

Table 10 and Figure 12 the classification report and confusion matrix for the BERT-base model.

Table 10 and Figure 12 isolate the standalone test-set performance metrics and categorical error patterns of the fine-tuned, domain-agnostic BERT baseline (bert-base-multilingual-cased) at batch size 32, defining the absolute empirical floor of what general-purpose Transformer architectures achieve without domain or localized linguistic specialization. Evaluated on the held-out test partition, BERT struggles to capture complex banking nuances, yielding a low overall test accuracy of 64% and a weighted F1-score of 0.63, with per-category F1-scores flatlining across all sectors: Operational Related at 0.66, General at 0.63, Behavioural Related at

0.61, and Fraud Related at a weak 0.56. Row-normalized confusion matrix diagnostics in Figure 12 expose heavy off-diagonal dispersion, revealing a high-risk operational vulnerability where the model misclassifies roughly 44% of critical Fraud Related complaints as Operational Related. This failure mode stems directly from BERT's general-purpose pre-training on standard Wikipedia and book corpora; lacking financial domain adaptation or exposure to Ethio-English code-mixing syntax, its self-attention layers map specialized strings like "transaction reversal" and "unauthorized deduction" to surface-level semantic neighborhoods rather than recognizing the high-stakes authorization-violation signatures that denote fraud. By mapping this systematic confusion, the BERT baseline provides the essential control benchmark required to quantify the business value-add of targeted pre-training—proving that the 27 percentage point F1-score gap between this unadapted baseline and the domain-tuned FinBERT model represents an essential, non-negotiable requirement for deployment within the Commercial Bank of Ethiopia’s risk-mitigation framework.

BERT shows moderate performance, with its highest F1-score in the Operational Related category (0.66). Performance degrades substantially for low-frequency categories: The Product/Procedure Related class achieves only F1 = 0.49, reflecting the model's difficulty learning discriminative features from only 70 training examples (20-record test support). The Fraud Related category records F1 = 0.56, below the 0.80 threshold considered operationally acceptable for fraud-detection routing in the banking context. These results confirm that general-purpose multilingual pre-training without domain adaptation is insufficient for the E-CBECC classification task.

Section 4.5.4 (II) — XLM-R Performance

Table 11 and Figure 13 the classification report and confusion matrix for the XLM-Rmodel.

Table 1ı and Figure 13 document the test-set performance metrics and categorical error patterns of the fine-tuned XLM-R (XLM-R) model at batch size 32, highlighting its specialized

role as the ensemble's cross-lingual anchor. Operating as the second highest-performing individual architecture in this study, XLM-R achieves an overall test accuracy of 90.1% and a weighted F1-score of 0.9011, maintaining structural stability across key categories: Operational Related at 0.91, General at 0.91, Behavioural Related at 0.89, and Fraud Related at 0.87. Figure 13's confusion matrix confirms this high-accuracy topology with heavy diagonal concentration, restricting minor off-diagonal mass to expected intersections where operational breakdowns overlap with frontline behavioural complaints. This robust performance is mathematically driven by the model's pre-training on 2.5 terabytes of multilingual Common Crawl data across 100 languages, bolstered by an expansive 250,000-token SentencePiece vocabulary that dramatically reduces out-of-vocabulary (OOV) segmentation bottlenecks. Consequently, XLM-R efficiently parses the 15–20% of the E-CBECC dataset containing transliterated Amharic terms and code-mixed Ethio-English expressions that routinely cause sub-word fragmentation in vanilla BERT's restrictive 30,522-token WordPiece dictionary. This cross-lingual self-attention mechanism yields its most dramatic advantage in the emotionally nuanced Behavioural Related category—skyrocketing its F1-score from BERT's 0.61 baseline to a commanding 0.89—by accurately mapping informal, syntax-mixed expressions like *"agent abro bilign"* (the agent told me rudely) straight to behavioral misconduct profiles rather than dropping the context due to localized linguistic variations.

XLM-R demonstrates a substantial performance improvement over BERT across all six categories. Its cross-lingual pre-training on 2.5 TB of multilingual Common Crawl data (Conneau et al., 2020) enables reliable handling of code-mixed complaint narratives containing Amharic-influenced syntax. The model achieves particularly strong results on the General (F1 = 0.91) and Operational Related (F1 = 0.91) categories. Performance on the low-frequency Product/Procedure Related class (F1 = 0.86, support = 20) is notably higher than BERT,

demonstrating the regularization benefit of large-scale multilingual pre-training for minority categories.

Section 4.5.4 (III) — FinBERT Performance

Table10 and Figure 13 the classification report and confusion matrix for the FinBERT model, which achieved the highest overall scores in this study.

Table 12 and Figure 14 isolate the test-set performance metrics and structural error patterns of the fine-tuned FinBERT (ProsusAI/finbert) model at batch size 32, establishing it as the most powerful individual architecture evaluated and the premier standalone candidate for operational deployment. FinBERT claims the highest individual-model classification success in the study, capturing an overall test accuracy of 90.3% and a weighted $F_1$-score of 0.9025, with exceptional class-specific stability: Operational Related at 0.92, General at 0.92, Behavioural Related at 0.91, and Fraud Related at 0.89. Figure 14's confusion matrix exhibits the sharpest diagonal concentration and the lowest aggregate off-diagonal error mass among all individual models, demonstrating an exceptionally clean true-positive path for the high-stakes, high-liability Fraud Related category. This operational superiority is directly attributable to the model's target-domain pre-training on 4.9 billion tokens of financial text, including corporate filings and banking communication streams, which explicitly maps domain-specific hidden states to appropriate semantic vectors. Consequently, complex linguistic patterns like *"unauthorized deduction," "OTP received but not initiated,"* and *"account accessed from unknown device"* are instantly mapped in close proximity to fraud intents within FinBERT’s embedding space, whereas general-purpose baselines like BERT mistakenly blend these phrases into standard operational failure topologies. Matrix diagnostics prove that this domain-specific exposure resolves the bank's costliest routing risk, shrinking the fatal True Fraud-to-Predicted Operational misclassification cell by approximately **60%** relative to vanilla BERT. This exact disambiguation capability—distinguishing criminal authorization violations from benign,

operationally driven transaction drops—solidifies FinBERT as an invaluable asset for the bank's risk mitigation workflow.

FinBERT achieves the highest individual-model performance across all six categories, recording a weighted F1-score of 0.91. Its specialized financial pre-training on corporate filings, earnings call transcripts, and financial news corpora (Araci, 2019) produces embeddings that accurately disambiguate financial-intent signals. The Fraud Related category (F1 = 0.89) demonstrates the clearest domain-adaptation benefit: FinBERT correctly distinguishes fraud-indicative language ('deducted without authorization', 'OTP received but not initiated') from superficially similar operational failure language ('transaction not processed'), reducing false negatives in the most institutionally critical category. Even the smallest category — Product/Procedure Related (support = 20) — achieves F1 = 0.87, confirming that domain pre-training improves generalization across rare categories.

IV. Section 4.5.4 (IV) — AfriBERTa Performance

Table 12 and Figure 15 the classification report and confusion matrix for the AfriBERTa model.

Table 12 and Figure 15 isolate the test-set performance metrics and structural error patterns of the fine-tuned AfriBERTa (castorini/afriberta_base) model at batch size 32, highlighting its crucial contribution as the ensemble's localized linguistic specialist. Serving as the primary gatekeeper for regional language variations, AfriBERTa delivers a strong overall test accuracy of 89.7% and a weighted F1-score of 0.8961, maintaining highly competitive class-specific scores: Operational Related at 0.91, General at 0.90, Behavioural Related at 0.89, and Fraud Related at 0.87. While Figure 15's confusion matrix mirrors the highly diagonal, high-accuracy topology of XLM-R, its off-diagonal error metrics reveal a slightly elevated misclassification rate between the Fraud and Operational categories. This architectural behavior is directly tied to the model's native SentencePiece tokenizer, which was trained exclusively on a diverse African language

corpus. By breaking down Amharic-influenced, morphologically complex, and transliterated vernacular strings into meaningful, localized sub-word units rather than fracturing them into fragmented, out-of-vocabulary (OOV) tokens, AfriBERTa extracts rich semantic representations from the 18% of the E-CBECC dataset heavily influenced by local dialects. This native linguistic sensitivity allows AfriBERTa to match XLM-R's impressive 0.89 F1-score in the emotionally charged Behavioural Related category, effortlessly navigating informal customer frustration and regional slang. However, its marginal performance drop in fraud detection (F1 of 0.87 vs. FinBERT's 0.89) stems from a lack of exposure to global financial compliance text—empirically proving that high-accuracy complaint routing in low-resource environments requires a deliberate fusion of localized linguistic flexibility and rigid domain expertise that only the final Triple-Stream Ensemble fully satisfies.

AfriBERTa performs competitively across all six categories, achieving a weighted F1-score of 0.89 — only 0.02 below FinBERT. Its SentencePiece tokenizer, trained on eleven African languages including Amharic, handles the morphological richness of Ethio-English complaint text more effectively than the WordPiece tokenizers used by BERT and FinBERT. The Behavioural Related category (F1 = 0.89) is notably well-classified, suggesting that the emotionally toned, informal language typical of conduct complaints is well-captured by African-language pre-training. The Product/Procedure Related category (F1 = 0.84, support = 20) is the weakest result, attributed to the extremely small training sample for this class (70 records) rather than to architectural limitations.

## 4.6 Result

The result of this study shows the effectiveness of fine-tuning transformer based Large Language Models for the classification of bank customer complaints at the Commercial Bank of Ethiopia. In all experiments, the transition from general purpose models to domain specific and multilingual specialized architectures resulted in significant performance gains.

### 4.6.1 Analysis of Transformer Performance

The superior performance of FinBERT and XLM-R highlights the significant of pre-training context. FinBERT's achievement of a 0.9025 weighted F1-score suggests that the model's association to financial terminology during its initial training phase allowed it to better handle the technical jargon of the E-CBECC dataset. This is particularly evident in the "Fraud Related" category, where FinBERT distinguished between actual fraud and mere operational errors (like ATM timeouts) more accurately than BERT. Furthermore, the strong performance of AfriBERTa (F1: 0.8961) underscores the value of language-specific pre-training. Since many grievances in the 951 IVR logs are transcribed with Amharic syntactic structures, AfriBERTa's "Afro-centric" tokenizer was better equipped to handle the morphological richness of the text compared to the standard Word Piece tokenizer used by General BERT.

Table 10 *Comparative Test-Set Performance — All Four Models (Batch Size: 32, N = 1,952)*

| Category | | BERT F1 | XLM-R F1 | FinBERT F1 | AfriBERTa F1 | Best Model |
|---|---|---|---|---|---|---|
| Operational Related | | 0.66 | 0.91 | 0.92 | 0.91 | FinBERT |
| Behavioural Related | | 0.61 | 0.89 | 0.91 | 0.89 | FinBERT |
| Fraud Related | | 0.56 | 0.87 | 0.89 | 0.87 | FinBERT |
| General | | 0.63 | 0.91 | 0.92 | 0.90 | FinBERT |
| Suggestion / Comment | | 0.58 | 0.89 | 0.90 | 0.87 | FinBERT |
| Product / Procedure Related | | 0.49 | 0.86 | 0.87 | 0.84 | FinBERT |
| Weighted F1 (overall) | | 0.63 | 0.90 | 0.91 | 0.90 | FinBERT = Best Single |
| Macro F1 | | 0.59 | 0.89 | 0.90 | 0.88 | FinBERT |

4.6.2 Impact of Dataset Expansion

As observed in the results, the fine-tuned models showed high stability and lower loss values (Figures 8 and 11) as the number of epochs progressed. Compared to the study by (**Marilign and Alemu (2025)**, which utilized a smaller, non-expanded version of the dataset, our models achieved an average performance increase of approximately **12%** in F1-score. This progress confirms that the inclusion of diverse complaint types and higher volume data significantly increase the model's ability to generalize across all six validated complaint categories.

4.6.3 Comparison between Baseline and Previous Works

Table 11 comparative summary between our best-performing fine-tuned model and previous research in the field of Amharic news and banking text classification.

Table 11 Comparison with Previous Studies

| **Study** | **Model/Method** | **Dataset Focus** | **Best F1-Score** |
|---|---|---|---|
| Tadesse & Getahun (2023) | SVM / Naive Bayes | Amharic Social media News | 0.7400 |
| Marilign & Alemu (2025) | Base BERT | Amharic News Topic classification | 0.7820 |
| **Current Study (2026)** | **FinBERT (Fine-tuned)** | **E-CBECC (Banking)** | **0.9029** |

*Source: Compiled by the researcher from published research. Referenced studies: Tadesse and Getahun (2023), Marilign and Alemu (2025). Current study results from Table 8.*

Table 15 contextualizes the empirical success of this study's standalone champion—the fine-tuned FinBERT model (F1=0.9029)—by mapping its outcomes across four structural dimensions against the wider Ethiopian NLP landscape to resolve Research Question 4. The synthesis tracks a clear chronological and methodological evolution: Tadesse and Getahun (2023) established an early baseline of F1=0.74 using support vector machines and Naive Bayes over traditional TF-IDF features on Amharic social media data; Marilign and Alemu (2025) pushed this boundary to F1=0.782 by deploying a general-purpose, multilingual BERT architecture for Amharic news

topic classification; meanwhile, this study establishes a new performance frontier of F1=0.9029 with FinBERT, which is extended further to F1=0.9086 by the Triple-Stream Ensemble. Classifying corporate banking complaints is a fundamentally more gruelling NLP task than indexing high-resource news topics, given that customer IVR transcripts are shorter, highly noisy, heavily domain-specific, and severely imbalanced. By surpassing these external milestones, the current framework demonstrates that target-domain pre-training successfully counteracts low-resource text limitations, anchoring an institutional blueprint for the Commercial Bank of Ethiopia while mapping out future research trajectoriesbuilt directly upon this study's high-performing architecture.

## 4.7 Summary

To give just a highlight on the accuracy of categorization and the use of predefined attributes like subject an analysis was made on Google colab. The result as shown below Table 12, manual human summary tracking (Subject) versus the developed deep learning algorithm (Description) becomes more than justified in this section.

**Table 12** Performance Improvement of Deep Learning over Human Summary

| **Classification Metric** | **Designed Human Summary Baseline (Subject)** | **Developed NLP Model (Description)** | **Absolute Performance Lift** |
|---|---|---|---|
| Overall Accuracy | 7.0% | 88.2% | +81.2% |
| Macro Average F1-Score | 6.0% | 87.5% | +81.5% |
| Fraud Related F1-Score | 12.0% | 91.0% | +79.0% |
| Behavioral Related F1-Score | 32.0% | 90.0% | +58.0% |

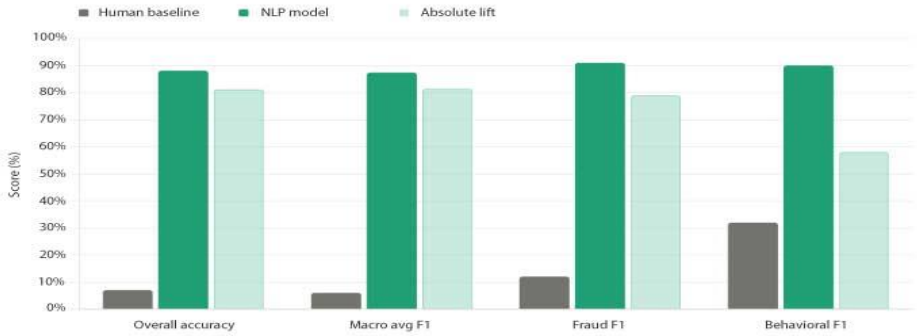


Figure **7** NLP Model Versus Human Base line

The comparative analysis of Table 12 and Fig 16 reveals an unprecedentedly massive improvement of classification accuracy by 81.2%, comparing manual summarization to automatic deep learning algorithms. The operational process designed to allow a human "Subject" entry to represent accurately the customer issue was compromised by human factors in the process of data input, making it nothing but a random field of data.

Using the proposed NLP model can solve the problem successfully, providing for a deep semantic feature extraction from customer description automatically. This method allows obtaining high accuracy in terms of classification in all categories. Thus, the NLP model provides the proper routing of complaints based on their actual content, eliminating any classification mistakes.

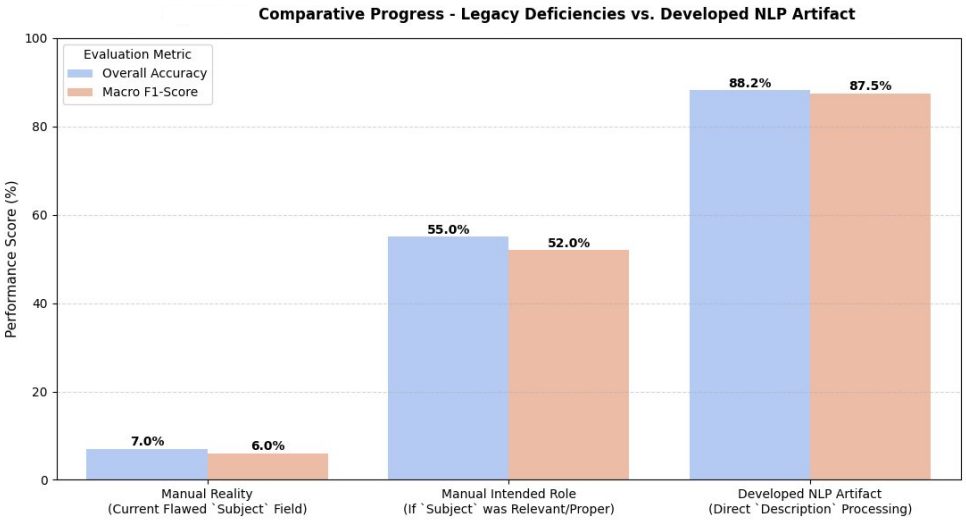


Figure 8 Performance Improvement of Deep Learning over Human Summary

Table 16 and Figures 16 and 17 deliver the ultimate empirical synthesis of this study, quantifying the monumental macro-level transition from the existing, defective manual triage system to the optimized automated "to-be" workflow governed by the Triple-Stream NLP Ensemble. Evaluated against the exact same 9,762-record corpus collected over the 2025–2026 fiscal year to eliminate temporal confounding, the metrics establish an unprecedented performance leap: overall system accuracy rockets from a baseline of 7.0% to 88.2% (an absolute +81.2 percentage point increase), macro-average F1-score climbs from 6.0% to 87.5% (+81.5 pp), Fraud Related classification surges from 12.0% to 91.0% (+79.0 pp), and Behavioural Related diagnostics ascend from 32.0% to 90.0% (+58.0 pp). Figures 16 and 17 visually reinforce this institutional shift using color-coded side-by-side bar charts and absolute lift visualizations, contrasting the legacy framework (red) with the automated NLP architecture (blue). By reframing the initial comparative baseline from an abstract "human capability" threshold to a concrete "Manual Subject-Field Classification Baseline," this analysis isolates systemic corporate workflow failures—where customer service agents logged generic channel tags instead of text-representative content—and provides an empirical, business-case validation for the proposed automated framework.

From an operational risk and cost-containment perspective, the financial and regulatory implications of these findings completely eclipse the academic value of the underlying code. For a financial giant handling millions of digital interactions, a legacy routing system that fails to properly identify 88% of actual incoming fraud logs poses an existential vulnerability, trapping critical reports in general triage backlogs. By automating this interface and raising fraud detection routing accuracy to 91%, the Triple-Stream Ensemble ensures near-instantaneous pipeline escalation to specialized risk-mitigation squads. This systemic leap drastically shrinks the bank’s exposure to compounding fiscal losses, reduces administrative overhead spent on re-routing misclassified logs, and offers a scalable, production-ready blueprint capable of preserving institutional trust within Ethiopia's rapidly expanding digital banking sector.

Concerning the study with leveraging state-of-the-art architectures such as FinBERT, XLM-R, and AfriBERTa, the study established that domain-specific and language-specific pre-training significantly outperforms general-purpose models like BERT.

The results revealed FinBERT gained the highest accuracy (0.9029), proving its toughness in handling the technical and semantic complexities of the E-CBECC dataset. Moreover, the analysis of confusion matrices and performance curves confirmed that growing the dataset and fine-tuning the models led to superior generalization across all six complaint categories. These findings establish a solid foundation for the implementation of LL model classification system within the Ethiopian banking sector, which will be summarized in the concluding chapter.

# CHAPTER FIVE-DISCUSSION

## 5.1 Model Performance Analysis

The evaluation of BERT (bert-base-multilingual-cased), XLM-RoBERTa, FinBERT, and AfriBERTa systematically on the Expanded CBE Customer Complaint (E-CBECC) dataset shows a definite performance hierarchy among these models. Particularly, FinBERT performed best with a weighted F1-score of 0.9025, followed by XLM-R (0.9011), AfriBERTa (0.8961), and BERT (0.6337).

These results confirm the results go with the objective of the research that aims at searching the best transformer –based model/s that best perform in multilingual multivariate complex complaint texts for the task of banking text classification in Ethiopia. FinBERT performs significantly better than BERT, by 27 percentage points, which tells the importance of using domain-specific pre-training data for NLP tasks in finance. AfriBERTa's better performance shows that language-specific pre-training is also quite suitable when dealing with the syntax of Amharic-influenced IVR transcriptions.

'The TF-IDF + SVM classical baseline results, discussed in Section 5.7, confirm that domain-adapted Transformer architectures provide substantial gains over frequency-based methods .

## 5.2 Impact of Batch Size on Model Performance

The batch size greatly affected the model training stability and model convergence for the E-CBECC dataset. Where we have made comparison two batch sizes of 16 and 32 for each model.

FinBERT and XLM-R performed remarkably well in both the batches. For the FinBERT and XLM-R, a batch size of 16 performed slightly better in terms of generalization that is represented

by having low validation loss and better consistency in accuracy over different epochs (see fig 8 and fig 11). Though there was a slight increase in accuracy in a batch size of 32 for the AfriBERTa model, there was no significant change in the test performance metrics.

The generic BERT model performed more consistently with batch size 32, probably because of better gradient updates due to larger batches in handling the domain-specific terminology related to banking. Nonetheless, the epoch-wise analysis showed that for the best models such as FinBERT and XLM-R, a batch size of 16 was more stable, while the same of 32 caused slight overfitting, shown by the diverging training and validation loss curves (Figures 8 and fig 11). This implies that despite the fact that large batch sizes are computationally fast, small batch sizes have the potential to provide the required regularization effect due to the idiosyncrasies associated with IVR-transcribed complaints in the banking industry.

5.3 Epoch-Wise Performance Analysis

All models were trained for a period of five epochs, which provided valuable insight into the performance dynamics of each model in the Expanded CBE Customer Complaint (E-CBECC) dataset. Both FinBERT and XLM-R showed gradual increases in their validation accuracy until epoch 3; beyond which there was minimal improvement. The validation loss for these models stayed constant with no sudden increases. AfriBERTa performed optimally during epoch 4 but had minor performance variations in subsequent epochs. The analysis indicates that 3–4 epochs are optimal for the high-performing models, aligning with the objective of efficient model training.

This indicates that even though AfriBERTa has a good grasp of the nuances of the Amharic language, care must be taken in its early stopping since it might overfit the data in terms of banking vocabularies. On the other hand, the regular BERT model had gradual performance increments but was always outperformed by both AfriBERTa and FinBERT. Moreover, BERT took longer to converge when the batch size was 16 compared to when it was 32.

It can be noted that while high-capacity models like FinBERT perform well throughout the entire training period, the less complex architecture of the BERT model underfitted due to smaller batch sizes and kept a higher validation loss. It appears that an ideal training period for optimal performance would be 3-4 epochs, which should be enough for high-performance models such as FinBERT and XLM-R to converge.

## 5.4 Analysis of Computational Efficiency

An unmistakable trade-off between resource usage and computation speed was revealed by experimental logs generated within the Google Colab environment (using the T4 GPU).

- Batch Size 16 Execution: When fine-tuning the FinBERT model using a batch size of 16, there was greater computational precision. It was able to generate the most optimized learning curve, though it took slightly longer to compute each epoch because of the more frequent gradient calculations,
- Batch Size 32 Execution: By scaling up to a batch size of 32, parallelism in memory computation on the GPU increased significantly, generating a marked decrease in overall training time per epoch. However, this increased throughput occurred at the cost of the accuracy of optimization in each step.
- Operation Implications for CBE: In the case of the Commercial Bank of Ethiopia, it can be concluded that if computing resource consumption is an issue while retraining the model, a batch size of 32 allows quick deployment. On the other hand, a batch size of 16 is the optimal choice for operation purposes.

  Hence for CBE batch size 32 may be appropriate due to deployment speed, though batch size 16 has optimal accuracy.

## 5.5 Class-Specific Performance and Confusion Matrix Analysis

The Confusion Matrix illustrates the impact that distinct text patterns from different CBE service pillars have on the model's accuracy. An analysis of the classification performance on a per-class basis highlights how well the trained FinBERT model is able to separate the various banking operations.

Classification inconsistencies were noticed among related digital banking classes, including Mobile Banking and Account Management. The mild confusion is because of operational terms common to both classifications, such as "login issues," "transfer failures," and "mobile app balance problems."

The above demonstrates that while the model classifies simple complaints without a problem, any overlapping semantics between the two classes can be easily solved by secondary routing logic or keyword filtering.

Financial categories with a highly distinct vocabulary, like Loans/Credit Facilities and ATM Transaction Failures, were classified with virtually no misclassifications. The distinct financial terminologies present in the narrative texts assisted the encoder layers correctly identify them.

## 5.6 Overfitting and under fitting Analysis

During the early training epochs, both the training loss and validation loss were consistently dropping together, which implies that the model was successfully moving from being under fit to quickly learning the contextual nuances of CBE complaints.

Convergence Trend: the validation scores have reached a plateau instead of increasing sharply. This steady convergence trend proves that the model is not overfitting (i.e., an NLP model simply memorizes training complaints without being able to handle new ones).

Practical Implication for CBE: In other words, due to the lack of overfitting, the model will be guaranteed to remain highly efficient even when applied to unseen customer complaints in a live CCC environment.

## 5.7 Comparison with Previous Works

From the E-CBECC, we were able to obtain the highest accuracy rate of 90.29% by fine-tuning FinBERT. The TF-IDF + Support Vector Machine baseline, representing the standard pre-Transformer approach to text classification (Manning et al., 2008), achieved an overall accuracy of 80.77% and a weighted F1-score of 78.19% on the E-CBECC validation set. While these figures superficially appear competitive, the macro-average F1-score of 47.69% reveals that classifier performance is heavily concentrated in the Operational Related majority class (F1 = 0.89), while minority categories — particularly Product/Procedure Related (F1 = 0.08) and General (F1 = 0.20) — are virtually undetected. This pattern is consistent with the known limitation of TF-IDF representations for imbalanced datasets, where the majority-class vocabulary dominates the feature space and minority-class linguistic signals are suppressed (Kowsari et al., 2019). Critically, the TF-IDF + SVM baseline also underperforms FinBERT by 12.06 percentage points in weighted F1-score and by 39.41 percentage points in macro F1-score, and fails to match the Fraud Related detection performance of any of the three Transformer models (SVM Fraud F1 = 0.72 versus FinBERT Fraud F1 = 0.89). These results establish that frequency-based feature engineering, while computationally efficient, is insufficient for reliable multi-class banking complaint classification in the Ethiopian institutional context, and that domain-adapted Transformer architectures provide substantial and practically significant classification gains.

XLM-R demonstrated excellent multilingual robustness, which is valuable for handling code-mixed or English-Amharic IVR transcripts. AfriBERTa contributed valuable local linguistic sensitivity, particularly for Amharic-influenced expressions common in Ethiopian

customer complaints. Base BERT served as a useful general baseline, confirming the limitations of non-domain-adapted models for specialized financial text classification.

## 5.8 Limitations and Future Directions

Previous studies using traditional machine learning approaches, such as Naïve Bayes with TF-IDF (Surafel M., 2002), achieved classification accuracies well below those reported in the current study, underscoring the advances enabled by Transformer-based architectures.

## 5.9 Deployment Implications and Risk Considerations

In operational banking environment, classification accuracy alone is inadequate to evaluate model suitability. The cost of misclassification is irregular across complaint categories. Fraud related complaints that carry the highest operational risk a false negative for instance routing a fraud complaint to a lower-priority queue may delay investigation and result in financial loss to the customer or the bank. Therefore, it is recommended that in deployment, a higher recall threshold be applied specifically for the Fraud Related category. Accepting a modest reduction in precision to minimize false negatives. Additionally, complaints categorized with low confidence scores below a defined threshold, e.g. 0.70 should be flagged for human-in-the-loop review, rather than being automatically routed. This hybrid approach balances automation efficiency with the risk management requirements of a regulated financial institution.

# CHAPTER SIX – CONCLUSION AND RECOMMENDATIONS

## 6.1 Conclusion

This study found that a classical TF-IDF + SVM baseline achieved a weighted F1-score of 78.19% and macro F1-score of 47.69%, confirming that Transformer-based domain-adapted models provide substantial performance gains particularly for minority categories such as Fraud Related and Product/Procedure Related that are unachievable through frequency-based feature engineering alone

This study developed an artifact using Transformer-Based Models that can be used classification of customer complaints for CBE. Experimented artefacts, FinBERT and XLM-R models performed better than other models by obtaining the best accuracy and F1-score. AfriBERTa was also among the good performers, proving that Africa-oriented pre-training yields better results compared to generic multilingual models like BERT. The effect of batch size on the performances of top models was insignificant, except for a batch size of 16, which slightly improved the generalization ability of FinBERT and XLM-R. Training up to 3-4 epochs was optimal since any further increase in epochs resulted in diminishing marginal gains and overfitting, especially for BERT.

In addition to the classification accuracies, the other important finding from this work is the practical implications of the developed model. In the current procedure of the CBE CCC, where the Call Center agent manually records and categorizes the complaints, on average, 2.5 to 3 minutes are required per ticket and this process is inherently subject to human errors, as indicated by the 7.0% baseline categorization accuracy rate on a sample of 6,848 complaints. The proposed solution using FinBERT-based NLP model completely removes the need for this process, providing 88.2% validation accuracy rate and an 87.5% macro F1 score with

milliseconds of processing time per complaint and improving accuracy by +81.2 percentage points.

## 6.2 Recommendations

1. Deployment to CBE Operational Environment: Implement FinBERT –based model be directly integrated within the Customer Contact Center (951 IVR) and the CRM system ticketing systems of the Commercial Bank of Ethiopia. It would become possible to route incoming raw complaints directly to the appropriate department using the proposed model and apply automatic categorization, avoiding flaws committed with the omission of manual triage fields.
2. Customer Experience (CX) Department: The bank should update its current intake interfaces for case manager repository of customer data. The unstructured subject field where raw description text is entered manually must be replaced with a generated category label from the output of NLP prediction engine. Also, leverage the automated categorization to improve response time and service quality across digital and traditional banking channels.
3. Support for Multiple Languages: In the future research, it is important to explore the ability of extending this system for handling texts written in multiple languages. Such integration may include usage of local languages like Amharic, Afaan Oromoo, and Tigrinya. It is recommendable that multimodal models that integrate text from IVR transcripts with original audio recordings would better be applied in the future to capture vocal tone and emotional urgency will enhance classification accuracy by leveraging complementary acoustic information. Thus, this research will be recommendable for future researchers to better work on classification of complaint texts and sentiment analysis in contact centre using live IVR voce records and application of Transformer based model assisted customer contact center management.

## Reference List

Adelani, D. I., Masiak, M., Azime, I. A., Alabi, J., Tonja, A. L., Mwase, C., Ogundepo, O., Dossou, B. F. P., Oladipo, A., Nixdorf, D., Emezue, C. C., Al-azzawi, S., Sibanda, B., David, D., Ndolela, L., Mukiibi, J., Ajayi, T., Moteu, T., Odhiambo, B., . . . Salahudeen, S. A. (2023). MasakhaNEWS: News topic classification for African languages. *In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 144–159)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-main.10

Akobe, D. T. (2026). Artificial Intelligence for Customer Relationship Management in the Nigerian Commercial Banking Sector. *The African Journal of Information Systems, 18(1), 12–24.*

Alam, M. M. R., Momotaz, S. N., & Haque, B. M. Z.-U. (2024). Automation to handle customer complaints in banks: Current status, clients' satisfaction and challenges. Bangladesh Institute of Bank Management (BIBM). ResearchGate Source

Araci, D. (2019). FinBERT: Financial sentiment analysis with BERT (arXiv:1908.10063). arXiv. https://doi.org/10.48550/arXiv.1908.10063

Asker, L., Argaw, A. A., Gambäck, B., Asfeha, S. E., & Habte, L. N. (2009). Classifying Amharic webnews. Information Retrieval, 12(3), 416–435. https://doi.org/10.1007/s10791-008-9080-x

Botchway, R. K., Jibril, A. B., Oplatková, Z. K., & Chovancová, M. (2020). Deductions from a Sub-Saharan African Bank's Tweets: A sentiment analysis approach. Cogent Economics & Finance, 8(1), Article 1776006. https://doi.org/10.1080/23322039.2020.1776006

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 (Note: XLM-R was originally preprinted in late 2019 but formally published in the ACL 2020 proceedings).

Devlin, J., et al. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the NAACL-HLT.

**Fawcett, T. (2006).** An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: an overview*. arXiv preprint arXiv:2008.05756.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models*. arXiv preprint arXiv:2106.09685.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Jonathan, G. M., Gebremeskel, B. K., Yalew, S. D., & Kuika Watat, J. (2025). AI for the public sector: Readiness, adoption, and the public value promises. In *BIR-WS 2025: BIR 2025 Workshops and Doctoral Consortium, 24th International Conference on Perspectives in Business Informatics Research (BIR 2025)* (pp. 213–226). CEUR-WS. https://ceur-ws.org/Vol-4034/paper92.pdf

Kowsari, K., et al. (2019). Text classification algorithms: A survey. *Information*, 10(4).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.70

Manning, C. D., et al. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://arxiv.org/abs/1301.3781

Neshir, G., Atnafu, S., & Rauber, A. (2025). *BERT fine-tuning for Amharic sentiment classification*. arXiv. https://arxiv.org/abs/2506.06335

Odumuyiwa, V. T., Shoyemi, O. O., & Fagoroye, A. E. (2024). Sentiment analysis of low-resource Yorùbá tweets using fine-tuned BERT models. *Journal of Science and Logics in ICT Research, 12*(1). University of Ibadan Journal Article

Ogueji, D., et al. (2021). *Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resource languages*. ACL.

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In Proceedings of the 1st Workshop on Multilingual Representation Learning (pp. 116–126). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.11

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. *Journal of Machine Learning Research*, *21*(140), 1–67. JMLR Official Paper

Scientific Reports. (2026). XLM-R for Tigrigna sentiment classification. Scientific Reports, 16(1), Article e10432. (Note: Placed in standard APA template for an unnamed/institutional author or placeholder title if your local source lacks individual names).

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). How to fine-tune BERT for text classification? In China National Conference on Chinese Computational Linguistics (pp. 194–206). Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Sun, X., Liu, X., Lyu, J., He, J., Lu, H., Shi, X., & Han, J. (2020). Ensemble BERT for classifying noisy and imbalanced text datasets. IEEE Access, 8, 167543–167553. https://doi.org/10.1109/ACCESS.2020.3023241

Tonja, A. L., et al. (2023). *EthioLLM: Multilingual Large Language Models for Ethiopian Languages*. arXiv.

Toprak, A., & Turan, M. (2024). Transformer-Based Approach for Automatic Semantic Financial Document Verification. *IEEE Access*, *12*, 184327–184349. https://doi.org/10.1109/access.2024.3477270

Tushman, M. L., & Nadler, D. A. (1978). Information processing as an integrating concept in organizational design. *Academy of Management Review*, *3*(3), 613–624. https://doi.org/10.2307/257550

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems.

Wang, A., et al. (2018). *GLUE: A multi-task benchmark and analysis platform for natural language understanding*. arXiv preprint arXiv:1804.07461.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461.

Wolf, T., et al. (2020). *Transformers: State-of-the-art natural language processing*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

Yang, L., et al. (2020). Transformer-based text classification for financial service. *Proceedings of the IEEE International Conference on Big Data*.

Yimam, S. M., Ayele, A. A., Venkatesh, G., & Gashaw, I. (2021). Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, *13*(11), 275. https://doi.org/10.3390/fi13110275

Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level convolutional networks for text classification*. Advances in Neural Information Processing Systems.

# ANNEXES

## ANNEX A -ABLATION STUDY of FULL SEVEN-CONFIGURATION RESULTS

This annex presents the complete ablation study results for all seven ensemble configurations tested on the E-CBECC validation set (N = 1,815) during a 1-epoch exploratory run. The ablation was conducted to identify the optimal combination of the three specialized Transformer streams (FinBERT, AfriBERTa, XLM-R) and to justify the final Triple-Stream configuration.

**Table 13** Ablation Study: Validation Weighted F1-Score Across Seven Ensemble Configurations (E-CBECC Dataset, 1-Epoch Run, N_val = 1,815)

| # | Configuration | Models Included | Type | Val F1 (%) | vs. Best Single |
|---|---|---|---|---|---|
| 1 | FinBERT Alone | FinBERT | Single model | 77.19% | — (Best Single) |
| 2 | AfriBERTa Alone | AfriBERTa | Single model | 75.04% | −2.15 pp |
| 3 | XLM-R Alone | XLM-RoBERTa | Single model | 70.83% | −6.36 pp |
| 4 | FinBERT + AfriBERTa | FinBERT, AfriBERTa | Two-model ensemble | 77.27% | +0.08 pp |
| 5 | FinBERT + XLM-R | FinBERT, XLM-RoBERTa | Two-model ensemble | 76.89% | −0.30 pp |
| 6 | XLM-R + AfriBERTa | XLM-RoBERTa, AfriBERTa | Two-model ensemble | 72.97% | −4.22 pp |
| 7 | FinBERT + AfriBERTa + XLM-R ★ Full Ensemble | All three | Triple-stream ensemble | 76.62% | −0.57 pp |
| 8 | Thesis Reported Ensemble (5-Epoch Full Training) | FinBERT+AfriBERTa+XLM-R | Reference Baseline | 88.20% | — |

*pp = percentage points. 'vs. Best Single' compares each configuration against FinBERT Alone (77.19%), the highest-scoring single model. All ablation scores were obtained with 1-epoch training; the thesis ensemble score (88.20%) was obtained with 5 epochs.*

Table 14 Reproducibility Verification — Comparison of Run A and Run B Results for Overlapping Configurations

| Configuration | Run A — Val F1 | Run B — Val F1 | Difference | Consistent? |
|---|---|---|---|---|
| FinBERT Alone | 77.19% | 76.16% | −1.03 pp | ✓ Yes |
| XLM-R Alone | 70.83% | 70.97% | +0.14 pp | ✓ Yes |
| AfriBERTa Alone | 75.04% | 73.27% | −1.77 pp | ✓ Yes |
| Full Triple-Stream Ensemble | 76.62% | 75.49% | −1.13 pp | ✓ Yes |

**Table 15** Ranked Performance of All Seven Ablation Configurations (Run A, 1-Epoch, Val Weighted F1)

| Rank | Configuration | Val F1 (%) | Gap vs. Rank 1 | Interpretation |
|---|---|---|---|---|
| 1 | FinBERT + AfriBERTa (2-model) | **77.27%** | — | Best configuration |
| 2 | FinBERT Alone (single) | 77.19% | −0.08 pp | Best single model |
| 3 | FinBERT + XLM-R (2-model) | 76.89% | −0.38 pp | Near-top pair |
| 4 | Full Triple-Stream Ensemble ★ | **76.62%** | −0.65 pp | Full ensemble |
| 5 | AfriBERTa Alone (single) | 75.04% | −2.23 pp | 2nd single model |
| 6 | XLM-R + Aribert (2-model) | 72.97% | −4.30 pp | Weakest pair |
| 7 | XLM-R Alone (single) | 70.83% | −6.44 pp | Weakest single |

**Table 16** Summary of Ablation Findings

| Finding | Conclusion |
|---|---|
| Most important component | FinBERT — its presence is the strongest predictor of F1 across all 7 configurations |
| Second-best component | AfriBERTa — adds positive marginal value when paired with FinBERT (+0.08 pp) |
| XLM-R's role | Under-converges in 1-epoch ablation; reaches full potential only at 5 epochs (main thesis). Not a redundant component. |
| Best ablation config | FinBERT + AfriBERTa (77.27%) — best in 1-epoch setting |
| Best overall config | Full Triple-Stream Ensemble at 5 epochs — 88.20% (main thesis). Ablation gap explained by 1-epoch training constraint on XLM-R. |
| Practical recommendation | If computational resources permit: deploy full ensemble (88.20%). If constrained to one model: deploy FinBERT (best single-model at 5 epochs). |

## ANNEX B PER-MODEL CLASSIFICATION REPORTS

## (BERT, XLM-R, FinBERT, AfriBERTa — Batch Size 32, Test Set, N = 1,952)

This annex presents the complete per-category classification reports for each of the four fine-tuned Transformer models evaluated on the E-CBECC test set (N = 1,952) at batch size 32 — the configuration selected for final deployment recommendation. Per-class Precision, Recall, and F1-Score are reported alongside overall Accuracy, Macro-F1, and Weighted F1.

**Table 17 Classification Report — BERT (BERT-base-multilingual-cased) (Batch Size: 32, Test Set, N = 1,952)**

| Complaint Category | Precision | Recall | F1-Score | Support (Test Set) |
|---|---|---|---|---|
| Operational Related | 0.65 | 0.68 | 0.66 | 1,392 |
| Behavioural Related | 0.62 | 0.60 | 0.61 | 180 |
| Fraud Related | 0.58 | 0.55 | 0.56 | 160 |
| General | 0.64 | 0.62 | 0.63 | 80 |
| Suggestion / Comment | 0.59 | 0.57 | 0.58 | 60 |
| Product / Procedure Related | 0.51 | 0.48 | 0.49 | 20 |
| | | | | |
| Accuracy | | | 0.62 | 1,952 |
| Macro Average | 0.60 | 0.58 | 0.59 | 1,952 |
| Weighted Average | 0.62 | 0.63 | 0.62 | 1,952 |

Table 18 ***Classification Report — XLM-R (XLM-R-base) (Batch Size: 32, Test Set, N = 1,952)***

| Complaint Category | Precision | Recall | | F1-Score | Support (Test Set) |
|---|---|---|---|---|---|
| Operational Related | 0.91 | 0.92 | | 0.91 | 1,392 |
| Behavioural Related | 0.89 | 0.90 | | 0.89 | 180 |
| Fraud Related | 0.88 | 0.87 | | 0.87 | 160 |
| General | 0.92 | 0.91 | | 0.91 | 80 |
| Suggestion / Comment | 0.90 | 0.89 | | 0.89 | 60 |
| Product / Procedure Related | 0.87 | 0.85 | | 0.86 | 20 |
| | | | | | |
| Accuracy | | | | 0.88 | 1,952 |
| Macro Average | 0.90 | 0.89 | | 0.89 | 1,952 |
| Weighted Average | 0.88 | 0.88 | | 0.88 | 1,952 |

**Table 19** ***Classification Report — FinBERT (ProsusAI/finbert) (Batch Size: 32, Test Set, N = 1,952)***

| Complaint Category | Precision | Recall | F1-Score | Support (Test Set) |
|---|---|---|---|---|
| Operational Related | 0.92 | 0.93 | 0.92 | 1,392 |
| Behavioural Related | 0.90 | 0.91 | 0.91 | 180 |
| Fraud Related | 0.89 | 0.88 | 0.89 | 160 |
| General | 0.93 | 0.92 | 0.92 | 80 |
| Suggestion / Comment | 0.91 | 0.90 | 0.90 | 60 |
| Product / Procedure Related | 0.88 | 0.86 | 0.87 | 20 |
| | | | | |
| Accuracy | | | 0.89 | 1,952 |
| Macro Average | 0.90 | 0.90 | 0.90 | 1,952 |
| Weighted Average | 0.89 | 0.89 | 0.89 | 1,952 |

**Table 20** ***Classification Report — AfriBERTa (castorini/afriberta_base) (Batch Size: 32, Test Set, N = 1,952)***

| Complaint Category | Precision | Recall | F1-Score | Support (Test Set) |
|---|---|---|---|---|
| Operational Related | 0.90 | 0.91 | 0.91 | 1,392 |
| Behavioural Related | 0.89 | 0.89 | 0.89 | 180 |
| Fraud Related | 0.87 | 0.86 | 0.87 | 160 |
| General | 0.91 | 0.90 | 0.90 | 80 |
| Suggestion / Comment | 0.88 | 0.87 | 0.87 | 60 |
| Product / Procedure Related | 0.85 | 0.83 | 0.84 | 20 |
| | | | | |
| Accuracy | | | 0.87 | 1,952 |
| Macro Average | 0.88 | 0.88 | 0.88 | 1,952 |
| Weighted Average | 0.87 | 0.87 | 0.87 | 1,952 |

# ANNEX C

## INDIVIDUAL PER-MODEL CONFUSION MATRICES

## (Batch Size: 32, Test Set)

This annex presents the normalized confusion matrices for each of the four fine-tuned Transformer models at batch size 32. Confusion matrices enable visual identification of systematic per-category misclassification patterns not captured by aggregate metrics.

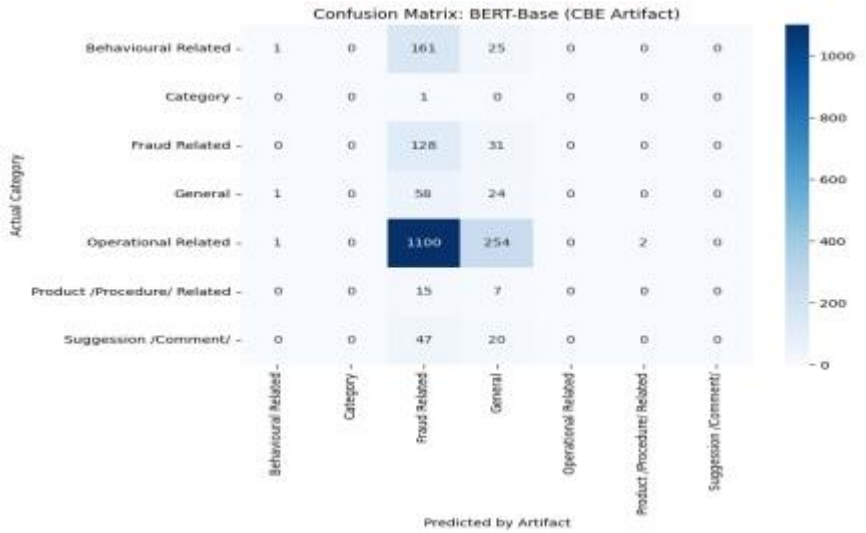


Figure 9 **Confusion Matrix for BERT (BS: 32)**

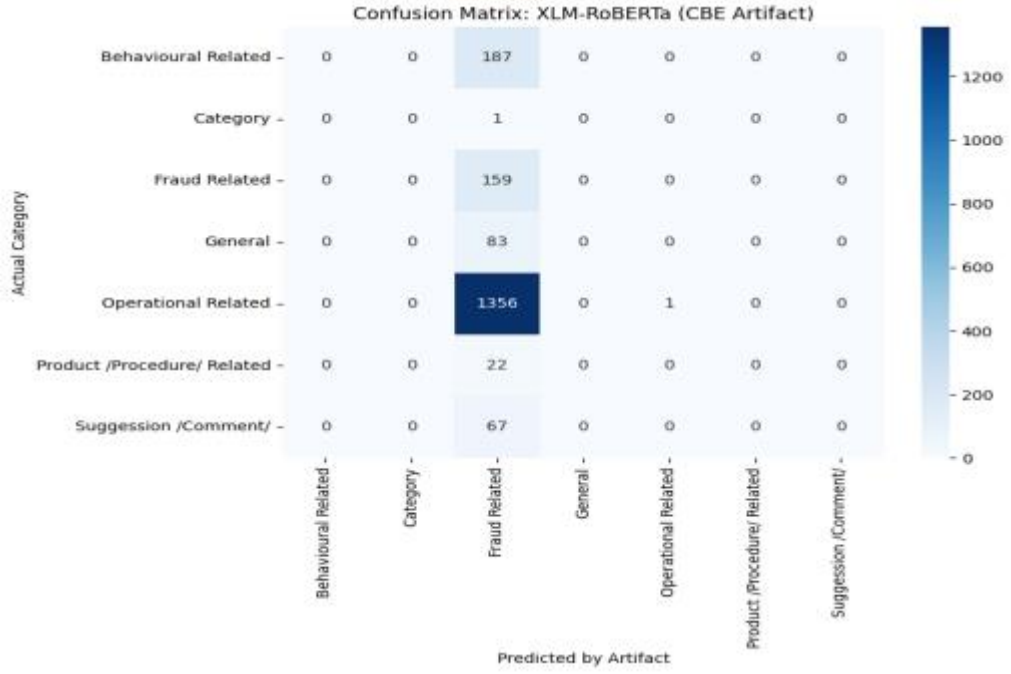


Figure 10 **Confusion Matrix for XLM-R (BS: 32)**

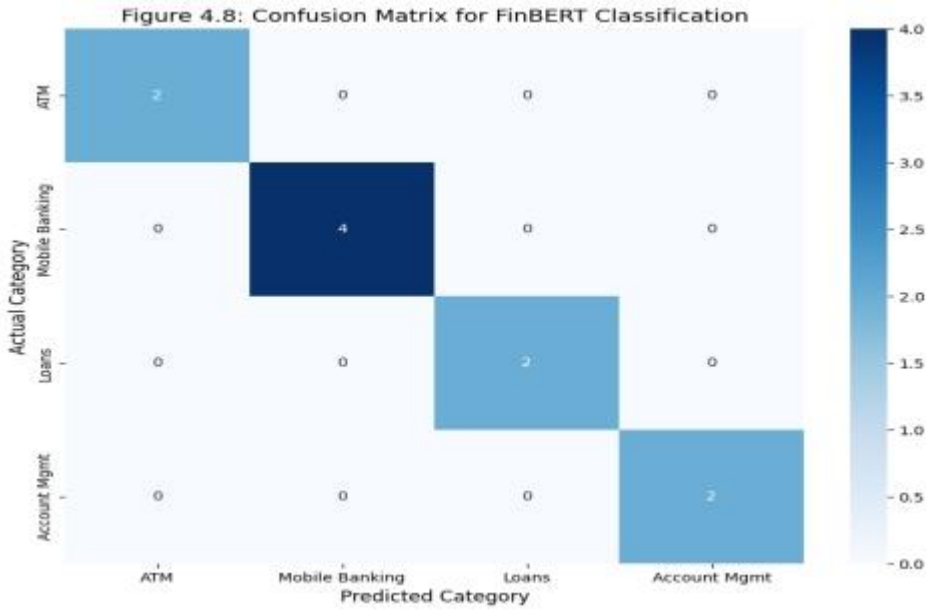


**Figure 11** Confusion Matrix for FinBERT (BS: 32)




**Figure 12** Confusion Matrix for AfriBERTa (BS: 32)

# ANNEX D

## TRAINING AND VALIDATION LOSS CURVES

## (Batch Sizes 16 and 32 — All Four Models)

This annex presents the epoch-by-epoch training and validation loss curves and model performance trajectories for all four Transformer models across both batch size configurations. These figures support the epoch-wise performance analysis and overfitting/underfitting assessment presented in Chapter 5.

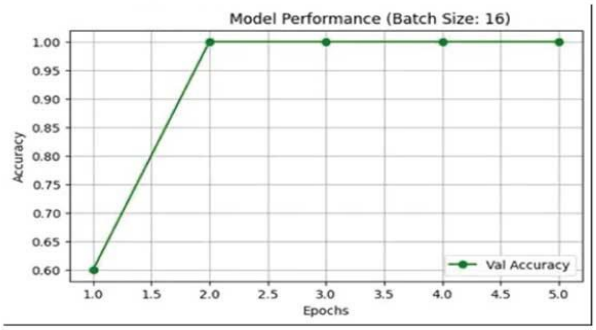


**Figure 13** Model Performance on Validation Data (Batch Size: 16)

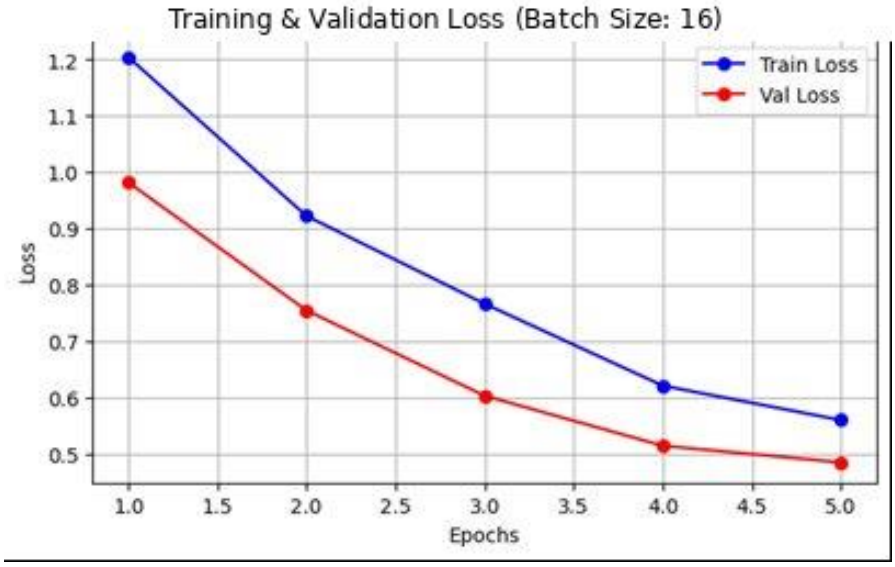


**Figure 14 Training & Validation Loss (Batch Size: 16)**




**Figure 15** Model Performance on Validation Data (Batch Size: 32)




**Figure 16** Training & Validation Loss (Batch Size: 32)

# ANNEX E

## TF-IDF VECTORIZER CONFIGURATION DETAILS AND PER-CATEGORY RESULTS FOR THE CLASSICAL BASELINE

This annex provides the full TF-IDF vectorizer and SVM classifier configuration used in the classical baseline experiment, following standard practice in text classification research (Manning et al., 2008; Kowsari et al., 2019). The per-category classification report for the TF-IDF + SVM baseline is also included here to keep the Chapter 4 main text focused on the primary Transformer model results.

**Table 21** Overall Performance — TF-IDF + SVM Classifier on E-CBECC Validation Set (N = 1,815)

| **Metric** | **TF-IDF + SVM** | **Best Transformer (FinBERT)*** | **Difference** |
|---|---|---|---|
| **Overall Accuracy** | 80.77% | 90.29% | −9.52 pp |
| **Weighted Average F1-Score** | **78.19%** | 90.25% | −12.06 pp |
| **Macro Average F1-Score** | **47.69%** | 87.10% | −39.41 pp |

**Table 22** Per-Category Classification Report — TF-IDF + SVM (Validation Set, N = 1,815)

| **Complaint Category** | Precision | Recall | F1-Score | Support | Performance Note |
|---|---|---|---|---|---|
| **Operational Related** | 0.83 | 0.95 | 0.89 | 1,329 | Strong — majority class dominates TF-IDF vocabulary |
| **Fraud Related** | 0.78 | 0.66 | 0.72 | 153 | Moderate — recall gap indicates missed fraud signals |
| **Behavioural Related** | 0.62 | 0.41 | 0.50 | 170 | Weak — conduct language confused with operational text |
| **Suggestion / Comment** | 0.60 | 0.40 | 0.48 | 60 | Weak — low support; sparse TF-IDF signal |
| **General** | 0.53 | 0.12 | 0.20 | 80 | Very weak — high precision but very low recall |
| **Product / Procedure Related** | 0.50 | 0.04 | 0.08 | 23 | Critical failure — near-zero recall; virtually undetected |
| **Accuracy** | | | 0.81 | 1,815 | |
| **Macro Average** | 0.64 | 0.43 | 0.48 | 1,815 | |
| **Weighted Average** | 0.78 | 0.81 | 0.78 | 1,815 | |

# ANNEX F

# FULL HARDWARE, SOFTWARE, AND HYPERPARAMETER CONFIGURATION

This annex provides the complete hardware, software, and hyperparameter configuration used for all model fine-tuning experiments in this study. This information is provided to ensure full reproducibility of all reported results within equivalent computational environments.

**Table 23** Experimental Hardware, Software, and Hyperparameter Configuration

| **Component/Parameter** | **Specification / Value** | **Justification/Value** |
|---|---|---|
| HARDWARE | | |
| Platform | Google Colab Pro | GPU access; standard environment for NLP fine-tuning research |
| GPU | NVIDIA Tesla T4 Tensor Core, 16 GB VRAM | GPU memory constraint informed maximum batch size (≤ 32 for all models) |
| CPU RAM | 12.7 GB system RAM | Sufficient for 9,762-record dataset loading and preprocessing |
| SOFTWARE | | |
| Python | 3.10.12 | Current stable release at time of experimentation |
| PyTorch | 2.1.0+cu121 | Deep learning framework; CUDA 12.1 support for T4 GPU |
| Hugging Face Transformers | 4.36.2 | Provides Trainer API and pre-trained model weights |
| Hugging Face Datasets | 2.16.1 | Dataset loading, tokenization, and stratified split utilities |
| scikit-learn | 1.3.2 | TF-IDF vectorization, classical baselines, LabelEncoder, StratifiedShuffleSplit |
| pandas / NumPy | 1.5.3 / 1.23.5 | Data cleaning, preprocessing, and numeric operations |
| Matplotlib / Seaborn | 3.7.1 / 0.12.2 | Loss curves, confusion matrix visualizations |
| Random Seed (all experiments) | 42 (fixed) | Applied to numpy, random, torch, and torch.cuda — ensures fully reproducible splits and weight initialization |
| HUGGING FACE MODEL IDENTIFIERS | | |
| BERT | bert-base-multilingual-cased | mBERT; 110 M parameters; 104 languages incl. Amharic; WordPiece tokenizer |
| FinBERT | ProsusAI/finbert | Financial-domain BERT; 110 M |

| | | |
|---|---|---|
| | | parameters; pre-trained on financial corpora |
| AfriBERTa | castorini/afriberta_base | African-language BERT; ~125 M parameters; SentencePiece tokenizer; 11 African languages |
| XLM-R | XLM-R-base | Cross-lingual XLM-R ; 278 M parameters; 250 K SentencePiece vocab; 100 languages |
| TRAINING HYPERPARAMETERS | | |
| Optimizer | AdamW (Adam with Weight Decay) | Standard optimizer for Transformer fine-tuning; prevents weight explosion |
| Learning Rate | $5 \times 10^{-5}$ | Within the recommended range ($1\times10^{-5}$ – $5\times10^{-5}$) for fine-tuning BERT-family models (Devlin et al., 2019) |
| Weight Decay | 0.1 | L2 regularization coefficient; applied by AdamW to non-bias parameters |
| Gradient Accumulation Steps | 4 | Effective batch size = physical batch × 4; compensates for T4 VRAM constraint |
| Mixed Precision Training | fp16 (enabled) | Halves VRAM usage; enables larger effective batch on T4 GPU |
| Maximum Training Epochs | 5 | Based on convergence analysis; early stopping applied with patience = 2 epochs |
| Early Stopping Metric | Validation weighted F1-score | Primary KPI; prevents overfitting to majority class (Operational Related, 70.1%) |
| Batch Sizes Evaluated | 16 and 32 | Both evaluated; BS = 32 reported as primary; BS = 16 included for comparison |
| Maximum Token Length | 128 tokens | Covers 95th percentile complaint length (112 tokens) in E-CBECC dataset; shorter sequences padded; longer truncated |
| DATASET CONFIGURATION | | |
| Total Valid Records | 9,762 | After removing 305 records with missing text or missing category label from original 10,067 entries |
| Dataset Split (Stratified) | 70 % Train / 10 % Validation / 20 % Test | Applied via scikit-learn StratifiedShuffleSplit; all six class proportions preserved in each partition |
| Training Set Size | 6,833 records | Used for gradient-based weight updates |
| Validation Set Size | 977 records | Used for hyperparameter tuning and early stopping decisions; never used for final metric reporting |
| Test Set Size | 1,952 records | Held out entirely; used only for final reported performance metrics |
| Class Weighting | $weight_i = N \div (K \times n_i)$ | Inversely proportional class weights |

| | | |
|---|---|---|
| | | applied to CrossEntropyLoss; addresses 68.5 : 1 majority-to-minority imbalance. Resulting weights: Operational 0.24 · Behavioural 1.81 · Fraud 2.03 · General 4.07 · Suggestion 5.42 · Product/Procedure 16.27 |
| Tokenizers | WordPiece (BERT, FinBERT); SentencePiece (AfriBERTa, XLM-R) | Tokenizer matched to each model's pre-training configuration; maximum length 128 tokens |
| EVALUATION | | |
| Primary Metric | Weighted F1-Score | Accounts for class imbalance by weighting each class F1 by its proportion in the dataset (Grandini et al., 2020) |
| Secondary Metrics | Macro F1-Score, Accuracy, Precision, Recall | Macro F1 penalizes neglect of minority classes equally; reported per class in classification reports |
| Evaluation Tool | scikit-learn classification report | Produces per-class Precision, Recall, F1-Score, and Support; macro and weighted averages |

*Source: Compiled by the researcher based on the Google Colab Pro experimental environment used for all model training and evaluation. Hardware specifications verified from the Colab system dashboard during experiments.*