



Arab American University
Faculty of Graduate Studies

**Analyzing Customer Reviews on Banks Mobile Applications:
A Bilingual Approach Using Natural Language Processing
for Topic Detection and Sentiment Analysis**

By

Aseel “Mohammed Osama” Maswadeh

Supervisor

Dr. Mohammed Maree

Co- Supervisor

Dr. Mujahed Eleyat

**This thesis was submitted in partial fulfillment of the
requirements for the Master's degree in
Data Science and Business Analytics.**

Feb/ 2026

© Arab American University –2026. All rights reserved.

Thesis Approval

Analyzing Customer Reviews on Banks Mobile Applications: A Bilingual Approach Approach Using Natural Language Processing for Topic Detection and Sentiment Analysis

By

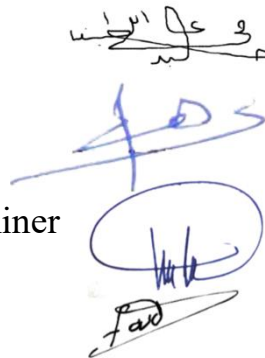
Aseel “Mohammed Osama” Maswadeh

This thesis was defended successfully on 14.2.2026 and approved by:

Committee members

Signature

1. Dr. Mohammed Maree: Supervisor
2. Dr. Mujahed Eleyat: Co-Supervisor
3. Dr. Mohammad Hamarsheh: Internal Examiner
4. Dr. Fadi Draidí: External Examiner



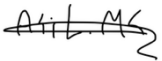
The image shows four handwritten signatures in blue ink, arranged vertically. The first signature is for Dr. Mohammed Maree, the second for Dr. Mujahed Eleyat, the third for Dr. Mohammad Hamarsheh, and the fourth for Dr. Fadi Draidí. The signatures are written in a cursive style.

Declaration

I declare that this thesis is my own work and has been prepared independently. All sources used have been properly acknowledged. This thesis has not been submitted for any other degree or qualification.

Student Name: Aseel Mohammed Osama Maswadeh

Student ID: 202112564

Signature: 

Date of Submitting the Final Version of the Thesis: 14.4.2026

Dedication

This thesis is dedicated to my father and mother, my family, my husband Mohammed Rahab, and my husband's family, in appreciation of their support and encouragement throughout my academic journey. I also extend my sincere thanks to everyone who contributed, supported, or assisted me during the completion of this work.

Acknowledgments

I would like to express my sincere appreciation to my supervisors, Dr. Mohammed Maree and Dr. Mujahed Eleyat, for their guidance, support, and valuable feedback throughout the course of this research. Their academic supervision and constructive comments greatly contributed to the completion of this thesis.

Abstract

Mobile banking applications have become a central channel for financial interaction, service delivery, and customer engagement in the Arab world due to the rapid expansion of digital banking services. This study aims to analyze Arabic user sentiment toward mobile banking applications and to identify the main topics that influence user satisfaction and dissatisfaction, in order to support service improvement and decision-making in the banking sector.

The study was conducted in 2025 using user reviews collected from Google Play for mobile banking applications affiliated with the top 100 Arab banks listed by The Banker (2024). Data were collected using a Python-based data extraction tool, resulting in more than 361,000 reviews. The dataset was filtered to include Arabic-language reviews only, while preserving the natural country distribution across the MENA region, reflecting real usage patterns and dialectal diversity. This dataset represents the study population, from which a labeled subset was used as the analytical sample.

The methodology involved developing a domain-specific Arabic preprocessing pipeline that includes text normalization, dialect handling, negation preservation, and meaningful bigram construction. Several annotation approaches were examined and compared, including manual labeling, rating-based sentiment inference, and GPT-assisted classification. Based on reliability and consistency, manual labeling was adopted as the final reference annotation. Class imbalance was addressed using class weights, SMOTE, and random oversampling. Sentiment classification was evaluated using traditional machine-learning models, an LSTM model, and transformer-based Arabic language models (AraBERT and MARBERT). Topic identification was conducted using a dual strategy combining probabilistic Latent Dirichlet Allocation (LDA) and GPT-based topic generation.

The results show that transformer-based Arabic models achieve the highest performance, with MARBERT outperforming all baselines due to its strong ability to capture dialectal and contextual information. Topic analysis indicates that positive reviews mainly emphasize ease of use, convenience, and successful transactions, whereas negative reviews are dominated by reliability-related issues such as login failures, unstable updates, slow performance, and application crashes. The study recommends prioritizing application reliability, strengthening pre-release testing, and adopting continuous sentiment monitoring to improve mobile banking services.

Keywords: Arabic Sentiment Analysis; Mobile Banking; Large Language Models; MARBERT; Topic Detection.

Table of Contents

Thesis Approval	I
Declaration	II
Dedication	III
Acknowledgments	IV
Abstract	V
Table of Contents	VI
List of Tables	IX
List of Figures	X
List of Abbreviation	XI
1. Introduction	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Research Aims	3
1.4 Research Questions	4
1.5 Research Objectives	4
1.6 Thesis Organization	5
2. Literature Review	7
2.1 Customer Experience – Definition	7
2.2 Social Media	9
2.3 Sentiment Analysis	10
2.4 Natural Language Processing	11
2.5 Topic Modeling for Arabic Banking Reviews	13
2.5.1 Latent Dirichlet Allocation (LDA)	14
2.5.2 GPT-4o-based Topic Modeling	15
2.6 Machine Learning	15
2.6.1 Random Forest	16
2.6.2 Naïve Bayes	17
2.6.3 Logistic Regression	18
2.6.4 Support Vector Machines (SVM)	18

2.7	Deep Learning.....	19
2.8	Large Language Models (LLMs) for Arabic NLP	20
2.8.1	Bidirectional Encoder Representations from Transformers (BERT)	21
2.8.2	AraBERT.....	23
2.8.3	MARBERT	24
2.9	Related Works	24
2.9.1	Sentiment Analysis Using Machine Learning.....	25
2.9.2	Sentiment Analysis with Deep Learning.....	27
2.9.3	Sentiment Analysis for Arabic Language	30
3.	Methodology.....	33
3.1	Introduction.....	33
3.2	Proposed Model for Arabic Sentiment analysis of Banking Application Reviews	33
3.2.1	Data Scraping.....	35
3.2.2	Data Preprocessing.....	35
3.2.3	Sample Selection.....	38
3.2.4	Data Labeling.....	40
3.2.5	Data Balancing.....	41
3.2.6	Term Frequency–Inverse Document Frequency (TF–IDF)	45
3.2.7	Machine Learning Models for Sentiment Classification	46
3.2.8	LSTM for Sentiment Classification	49
3.2.9	AraBERT for Sentiment Classification.....	51
3.2.10	MARBERT for Sentiment Classification.....	52
3.2.11	Topic Modeling.....	54
3.3	Conclusion	57
4.	Results and Discussion	58
4.1	Introduction.....	58
4.2	Experimental Setup.....	58
4.2.1	Implementation Environment	58
4.2.2	Libraries and Tools.....	59
4.2.3	Experimental Evaluation Strategy	59
4.2.4	Evaluation Metrics	60
4.3	Evaluation of Annotation Consistency Across Labeling Techniques	62

4.4	Evaluation Model.....	63
4.4.1	Evaluation of Machine Learning Models	63
4.4.2	Evaluation of Deep Learning Models	68
4.4.3	Evaluation of LLM Models	77
4.5	Topic Modeling and Semantic Interpretation of User Reviews	85
4.6	Conclusion	98
5.	Conclusion and Future Work	100
5.1	Conclusion	100
5.2	Future Work	103
	References	104
	الملخص	104

List of Tables

Table 3.1: Examples of Arabic Text Before and After Each Preprocessing Step	37
Table 3.2: LSTM Model Architecture and Parameters	50
Table 4.1: Agreement and Difference Rates Between Sentiment Labeling Techniques	63
Table 4.2: Machine Learning Performance Under Different Balancing Techniques	65
Table 4.3: Machine Learning Cross-Validation Results Under Different Balancing Techniques	67
Table 4.4: Classification Report of the LSTM Model Trained with SMOTE for Binary Arabic Sentiment Classification	70
Table 4.5: Classification Report of the LSTM Model Trained with Random Oversampling for Arabic Sentiment Classification	74
Table 4.6: Training and Validation Performance of AraBERT	78
Table 4.7: Classification Report of AraBERT Model	79
Table 4.8: Training and Validation Performance of MarBERT	82
Table 4.9: Classification Report of MarBERT Model	83
Table 4.10: Summary of Preprocessed Corpora Used for LDA Topic Modeling	85
Table 4.11: Selected Number of Topics and Coherence Scores for LDA Models	87
Table 4.12: Most Frequent GPT Topics	93
Table 4.13: Shared Topics Across Positive and Negative Reviews	94
Table 4.14: Comparison Between LDA and GPT-Based Topic Modeling	97

List of Figures

Figure 2.1: Random Forest architecture [55].....	17
Figure 2.2: Support vector machines architecture [63].....	19
Figure 2.3: Long Short-Term Memory architecture [67].....	20
Figure 2.4: Transformer architecture [73].....	22
Figure 2.5: BERT architecture [71].....	23
Figure 3.1: The Proposed Framework for Arabic Sentiment Analysis of Banking Application Reviews.....	34
Figure 3.2: Dataset collected from Google Play.....	36
Figure 3.3: Dataset After Preprocessing	38
Figure 3.4: Rating Distribution of Mobile Banking Application Reviews Before Sampling.....	39
Figure 3.5: Class Distribution Before and After Applying Class Weights.....	43
Figure 3.6: Class Distribution Before and After Applying SMOTE.....	44
Figure 3.7: Class Distribution Before and After Applying Oversampling.....	44
Figure 3.8: Class Distribution Before and After Applying Undersampling.....	45
Figure 4.1: Training and Validation Accuracy of LSTM	69
Figure 4.2: Confusion Matrix of the LSTM Model with SMOTE	71
Figure 4.3: ROC Curve of the LSTM Model with SMOTE	72
Figure 4.4: Training and Validation Accuracy Using Random Oversampling	73
Figure 4.5: Confusion Matrix of the LSTM Model for Binary Sentiment Classification using Random Oversampling	75
Figure 4.6: ROC Curve of the LSTM Model for Binary Sentiment Classification using Random Oversampling.....	76
Figure 4.7: Confusion Matrix of AraBERT Model for Binary Sentiment Classification	80
Figure 4.8: Confusion Matrix of MarBERT Model for Binary Sentiment Classification	84
Figure 4.9: Coherence Score Comparison for LDA Topic Models by Number of Topics.....	86
Figure 4.10: Topic Distribution in Positive Reviews Using LDA	88
Figure 4.11: Topic Distribution in Negative Reviews Using LDA.....	89
Figure 4.12: Word Cloud Visualization by Sentiment	90
Figure 4.13: LDAvis Interactive Visualization of Positive Topics in Arabic Banking Reviews ..	91
Figure 4.14: LDAvis Interactive Visualization of Negative Topics in Arabic Banking Reviews.	92
Figure 4.15: Top Frequent Positive and Negative Topics Identified by GPT-Based Topic Modeling.....	95

List of Abbreviation

Abbreviation	Full Term
NLP	Natural Language Processing
SA	Sentiment Analysis
ML	Machine Learning
DL	Deep Learning
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LLMs	Large Language Models
GPT	Generative Pre-trained Transformer
GPT-4o	GPT-4 Omnimodel
BERT	Bidirectional Encoder Representations from Transformers
mBERT	Multilingual BERT
AraBERT	Arabic BERT
MARBERT	Social-Media Arabic BERT
TF-IDF	Term Frequency–Inverse Document Frequency
SVM	Support Vector Machine
NB	Naïve Bayes
LR	Logistic Regression
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling Technique

Abbreviation	Full Term
LDA	Latent Dirichlet Allocation
API	Application Programming Interface
NER	Named Entity Recognition
POS	Part of Speech
MSA	Modern Standard Arabic
MRMR	Minimum Redundancy Maximum Relevance
HDFS	Hadoop Distributed File System
DLMNN	Deep Learning Modified Neural Network
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
MLP	Multi-layer Perceptron
KNN	K-Nearest Neighbor
BoW	Bag of Words
GloVe	Global Vectors for Word Representation
ReLU	Rectified Linear Unit
MENA	Middle East and North Africa
COVID-19	Coronavirus Disease 2019

1. Introduction

1.1 Background

The growth of digital technologies, along with the popularity of the usage of mobile applications, has resulted in the generation of colossal amount of user-created content on online platforms. Today, these applications have become major platforms on which people air their views, rate services, as well as provide real-time experiences. This tendency increased in the COVID-19 crisis when the transition to digital services compelled customers and financial institutions to heavily use mobile banking as their daily activity for money transfer and financial transactions [1], [2]. Consequently, the worldwide usage of the internet increased, and the number of social media users amounted to 5.24 billion as of February 2025 [3].

This booming expansion has two significant deliverables. In the first place, it opens issues of privacy, data security, and handling of huge, unstructured textual data. Conversely, it provides good opportunities of using sentiment analysis (SA) to draw conclusions about user behavior and online communication trends [4]. These issues are especially valuable in the banking industry where the perception of the customer matters a lot in terms of trusting an institution. Bad word or dissatisfaction can quickly kill off confidence, create regulatory issues and inflict financial harm. Thus, the necessity to constantly measure and analyze user sentiment has turned into a strategic need of current businesses.

SA is now a field of focus within natural language processing, where it tries to identify and determine the tone of emotion or the evaluative attitude that resides within a written communication [5]. In addition to its computational basis, it has emerged as a useful decision support system for organizations that are interested in knowing what customers think about their services and products in order to respond actively to their needs. Studies have emphasized the fact that the experience gained by monitoring the feedback of users plays a major role in strategic decisions, be it in creating new products or improving on the services, or services that constantly receive negative reviews [6].

Over the years, various methodological models have been created to help in the SA. These are the lexicon-based methods, ML models, DL Models, and hybrid systems that combine the lexicon-based techniques and ML-based models to achieve synergistic advantages. One of the oldest and most interpretable approaches is the lexicon-based method, which relies on the predefined dictionaries of sentiment, in which the words are linked with the value of polarity or the degree of intensity, allowing one to effortlessly rate the sentiment of text [7]. On the contrary, ML-based approaches are based on statistical learning, and they can be classified as supervised, unsupervised, or semi-supervised models, depending on the availability of annotated data [8]. Algorithms used in supervised learning are trained on labeled data to predict opinions (e.g. positive, neutral, or negative) whereas unsupervised models do not require annotated data and therefore are applicable where labeling is expensive or unavailable. Semi-supervised learning, a small annotated dataset guides the learning process, while the abundant unlabeled data helps the model generalize better by capturing the underlying structure of the data distribution. Collectively, the set of techniques can be viewed as a powerful toolkit of analysis that can reveal important trends in large amounts of user-generated material in many fields. Add some discussion on DL models.

In light of the rapid digital transformation of financial services and the strategic importance of understanding customer perceptions, it becomes evident that advanced and context-aware analytical approaches are required to effectively process Arabic mobile banking reviews. While existing sentiment analysis techniques provide valuable insights, their applicability in complex, dialect-rich, and domain-specific environments such as Arabic financial services remains limited. Therefore, a focused investigation into robust sentiment classification and topic extraction methodologies tailored to this context is necessary to bridge the gap between theoretical advancements and practical deployment in real-world banking systems.

1.2 Problem Statement

Although the world is rapidly changing to digital financial services, and mobile banking is gaining popularity in the Arab world, users' opinions expressed as online reviews are not well understood. The complexity of the Arabic language remains one of the key issues due to its rich morphology, extensive inflection, sporadic spelling, and varying dialects provide a unique challenge to the accurate SA. The literature in this context remains at a small scale, with the research typically

using small or domain irrelevant data, imbalanced annotation criteria and excessively focused on Modern Standard Arabic, which is not representative of the informal and dialect-heavy language typically used in the real comments that users actually post.

Mobile banking reviews have a spontaneous and colloquial style which introduces further challenges to the traditional machine learning methods, whereas deep learning models like LSTM need large and balanced labelled datasets which are uncommon in Arabic financial settings. Even though the Arabic Large Language Models such as AraBERT and MARBERT have wider language coverage, there are limited studies on how they perform in the financial sector. Topic extraction is also essential to understand the concerns of the users, however, classical models such as LDA are not suitable to deal with short documents that are noisy and have dialects, as is the case with online reviews.

1.3 Research Aims

In this research project, we aim to develop a solid and linguistically conscious sentiment and topic analysis framework that can be used for analyzing the sentiment orientation of online customer reviews on mobile banking applications that are expressed in Arabic language. The research is dedicated to the discussion of the particular issues of Arabic text, such as complicated morphology, a variety of dialects, and sound of financial content created by users. We aim at developing and testing a full analytical pipeline with the involvement of various annotation strategies and diverse modeling methods. These methods include classical ML algorithms like the SVM, Logistic Regression, Naive Bayes, and the Random Forest, and DL models that utilize neural networks like Long Short-Term Memory (LSTM); in addition to the exploitation of high-performing Arabic Large Language Models, such as the AraBERT and MARBERT. In addition, probabilistic topic modeling methods and semantic models based on Large Language Models will be examined. Finally, the thesis aims to provide scalable, precise, and domain-specific sentiment-topic analysis system that might assist financial institutions to comprehend customers opinion better, improve their digital banking authenticate, and increase user satisfaction within the Arab banking environment.

1.4 Research Questions

In order to outline the orientation of this thesis and justify its methodological decisions, a list of research questions was generated. The questions aim to address the primary issues that have been identified in the Arabic sentiment and topic analysis and give a clear roadmap towards achieving the research goals:

- How can a unified sentiment–topic analysis framework be designed to efficiently handle large-scale Arabic mobile banking reviews while overcoming linguistic obstacles such as complex morphology, orthographic inconsistency, and dialectal diversity?
- What impact do different annotation strategies – including manual labeling, rating-based labeling, and GPT-4o-assisted annotation – have on the reliability, consistency, and overall quality of sentiment classification in Arabic user-generated content?
- Which ML classifiers (SVM, Logistic Regression, Naive Bayes, Random Forest) and deep learning models (LSTM) provide the most accurate and resilient sentiment predictions when applied to Arabic banking reviews, especially under varying data-balancing and evaluation settings?
- How do Arabic Large Language Models such as AraBERT and MARBERT perform compared to traditional ML and deep learning approaches in handling context, dialectal differences, and sentiment ambiguity in noisy financial text?
- What distinctions can be observed in performance, interpretability, and topic coherence between classical topic modeling techniques like Latent Dirichlet Allocation (LDA) and more advanced semantic approaches such as GPT-4o when extracting themes from Arabic mobile banking application reviews?

1.5 Research Objectives

The primary aim of this thesis is to develop a high-quality and consistent sentiment and topic analysis pipeline for analyzing Arabic mobile banking reviews, with the help of cutting-edge ML, deep learning, and large language models. Since Arabic is a linguistically complicated language, and a range of dialects are involved in user-generated content, the proposed work is expected to

address the most common issues of dialect variation, inconsistent quality of annotations, or a class imbalance, and to retain the interpretability and high predictive accuracy at all the levels of analysis.

The project is aimed at the construction of end-to-end analysis pipeline. It starts with a large-scale preprocessing and normalization of Arabic text and proceeds with a multi-strategy annotation process with a combination of manual labeling, rating-based labeling, and GPT-4o-assisted annotations to enhance reliability and ground-truth fidelity. To classify sentiments, we compare a variety of models, both traditional ML (Support Vector Machine, Logistic Regression, Naive Bayes and Random Forest) and neural networks (Long Short-Term Memory (LSTM) networks and state-of-the-art Arabic Large Language Models (AraBERT and MARBERT) that can better represent contextual and dialect-specific information.

In addition to sentiment detection, the study also aims at discovering latent topics in banking reviews through both probabilistic topic modeling Latent Dirichlet Allocation (LDA) and semantic topic extraction with GPT-4o to create more meaningful and understandable topic structures. The integration of these complementary methods into a single analytical framework provides a scalable, precise, and linguistically strong set of tools that can assist financial institutions to comprehend user concerns better, improve the quality of digital banking, and customer experience within the Arab banking sector.

1.6 Thesis Organization

This thesis has been structured into five key chapters so as to make a systematic and coherent presentation of the research.

Chapter One presents the overall setting of the research by defining the significance of the analysis of Arabic customer reviews in the mobile banking industry, then indicating the statement of the problem, research aim, guiding research questions, and objectives of the research.

Chapter Two presents the literature review which gives the theoretical basis of the study. It starts with the introduction to the topic of customer experience and social media role, covers

sentiment analysis, topic modeling approaches with LDA and GPT-based techniques, proceeds to machine learning and deep learning, and finally finishes with studies on sentiment analysis, both in general and Arabic.

Chapter Three introduces the proposed methodology and overall analytical pipeline, data collection procedure, Arabic text preprocessing strategies, annotation techniques, data balancing techniques, feature extraction, sentiment classification models such as traditional machine learning, LSTM, AraBERT, and MARBERT, and dual topic modeling approach using LDA and GPT.

Chapter Four describes the experimental design and gives an in-depth discussion of the findings, such as model analysis, cross-analysis of the findings between the different annotation and balancing techniques, sentiment analysis results, and interpretation of topics that were identified in the reviews submitted by bi-lingual Arabs.

Chapter Five summarizes the key findings, contributions of the research, and future directions that will further the analysis of Arabic sentiment and topic in mobile banking and greater application of digital finance.

2. Literature Review

This chapter gives a coherent overview of theoretical basis, as well as practical frameworks pertaining to our study. It starts with a general customer experience introduction, and proceeds to analyzing the influence of social media in consumer communication. The next portion of the discussion is about the text analysis of sentiment and natural language processing, which are the important elements to understand the opinions of the users. The review has operated in the field of ML methods, then deep learning methods, and finally, the discussion of modern language modeling methods like those proposed by the BERT architecture. In order to have a clear background on SA contribution towards the improvement of customer experience, the chapter ends with a review of research on SA through both ML and deep learning, with studies devoted to the Arabic language.

2.1 Customer Experience – Definition

Customer experience is defined as a mix of the cognitive and emotional responses developed in the mind of the customer in response to an interaction with any given service regardless of whether the interaction is one on one or indirect [9] [10]. It has not been adequately explored, although its significance is increasingly becoming important especially in industries where competitive differentiation is highly needed like in the banking sector. Customers in the financial sector are faced with most diverse touch points of service such as accounts and cards, online and offline banking services, etc. and these touchpoints have a significant bearing on the choices that customers make when deciding whether to use them again or not and whether to be brand loyal or not. Research indicates that improved customer experience translates to improved customer satisfaction and loyalty subsequently increasing profitability of the financial institution [11] [12] [13]. Owing to this influence, banks are becoming more aware that they should learn, quantify, and enhance customer experience in order to maintain a competitive advantage in an ever-growing global marketplace.

Due to a number of studies, the factors that affect customer experience (CE) in the banking sector are discussed. Garg et al. (2014) emphasized the convenience as a fundamental factor in India, in

addition to customer interaction, employee behavior, speed of service, service overall quality, and digital capabilities [14]. As it was stressed by Chauhan et al. (2022), CE is defined by functional (quality, trust, and convenience), mechanical (website design, and usability), and human-related dimensions (particularly, the manner in which complaints are addressed) [15]. Ramaswamy et al. (2021) have determined security, network performance, and convenience to be significant contributors of CE [16]. Furthermore, Ashrafpour et al. (2021) observed that functional and emotional experience is very important in the context of Iran, and the argument presented by the authors is that both system and service quality are fundamental to the online banking experience [17]. In the context of mobile banking, Shaikh et al. (2020) discovered that user awareness, perceived usefulness, and ease of use were the main antecedents of CE [18]. According to Moliner-Tena et al. (2019), customer engagement and customer behavior in Spain were primarily influenced by customer engagement [19]. On the same note, Mbama and Ezepeue (2018) documented in the United Kingdom that the quality of services, perceived value, usability, employee-customer relationships, functional quality, and perceived risk, make a significant difference in CE in the banking sector [20].

According to Szmigin and Piacentini (2018), the attitude can be defined as a set of thoughts (beliefs and cognitions), emotions (feelings and affect) and behaviors (actions) [21]. They elaborate that attitude is comprised of three main elements: the affective element which is the emotional reaction of a person on an object; the behavioral element which is the actions by the attitude and the cognitive element which is the beliefs and perceptions a person has. These elements affect the important customer behaviors like repurchase, changing their provider or referring services, which are critical activities in the performance of a bank.

Psychologically, Oskamp and Schultz (2005) explain the attitudes as acquired responses, which can be rated as positive or negative and have qualities of strength, consistency, and stability [22]. According to their work, an attitude is a product of the association between particular object and value judgments that a person makes of that particular object.

The study has examined how a customer attitude can be influenced and evolving behavior of the customer in the banking industry. As a fact, Dauda and Lee (2015) note that a significant portion of customers has not been quick to use online services, whereas banks have been persuading

traditional branch customers to use digital platforms- a task that has proven to be difficult in a number of developing countries [22]. In line with this, Mwanajimba (2019) documented that Tanzanian customers showed negative attitudes towards banking technologies, which was mainly caused by the lack of awareness and cultural beliefs [24]. Reliability was found by Sakhaei et al. (2014) in Iran as the most affecting pillar that drives the acceptance of internet banking [25]. Similarly, Shrestha et al. (2020) established that the perceived usefulness, perceived risk, and perceived ease of use had a significant influence on the willingness to adopt online banking in Nepal, and that the overall user experience was generally positive [26].

Emotional relationship also is significant in determining attitude and behavior of customers towards banking services. This is why SA is a good method to quantify such attitudes and emotional attachment. The current study is unique in that it analyzes customer sentiment toward mobile banking services using user-generated reviews collected from Google Play, rather than relying on traditional surveys or general-purpose social media data. This method has a number of benefits, such as the ability to cover a myriad of attributes and opinions, be able to see the user experiences in real time, and being cheaper than other conventional research techniques like surveys or focus groups.

2.2 Social Media

Social media analytics (SMA) is an evolving area of research within the last few years, with numerous industries showing an increasingly strong interest in it as the number of digital platforms keeps growing [27] [28]. This growing interest is largely driven by the exponential increase in user-generated content and the strategic value of real-time consumer insights in highly competitive markets. SMA is concerned with the study of relationships, opinions and emotions of people through the role of internet interactions. The organizations are now able to interact with the customer directly and keep a close watch on the perception of the customers via their social media presence. The knowledge obtained on such platforms is applied in decision-making, business strategy, response to complaints and to gauge the attitude of the people towards the products, services and the brand image. In the banking sector in particular, where trust, reputation, and service reliability are critical success factors, understanding online customer discourse becomes

even more essential. This has spurred the banks to engage their customers online more and the need to conduct more research on the importance of maximizing on the use of social media.

Social networks like Facebook, Instagram, Twitter, and Google Play provide the end user with the platform allowing them to express their opinions on numerous issues with no restrictions. In 2004, Facebook was created and it became one of the most influential networks, but tough regulations of privacy do not allow using user data as a source of research because data availability is conditioned by user consent [29]. However, Twitter, which was launched on July 13, 2006, does not restrict public tweets but makes them available via its API although only within a seven-day window (as of the most recent limitations) [30]. This therefore means that meaningful understanding of consumer views can only be achieved when social talks are carefully listened and interpreted [28]. To do so, the current research paper utilizes SA methods to measure emotional polarity, identify trends in behavior, identify possible churn drivers, and get better insights into the experience customers go through when using digital banking services.

2.3 Sentiment Analysis

The success of any business or event, therefore, requires organizations to know the kind of opinions and attitudes that their target audience will have on social media. In the world of new technologies, social platforms provide an available and direct channel avenue where individuals express their emotions, responses, and opinions regarding different experiences freely and instantly. The messages left by the customer or the participants in the event are usually candid and offer good information in written format. Consequently, social media has turned out to be a significant resource in gauging the effectiveness of product promotion, events and brand success, and organizational operations.

These are platforms, which enable analysts to understand the sentiment of the people concerning an event or a company, and, therefore, can be a bountiful source of information when making strategic decisions. In order to measure success accurately, one should have a structured approach- this is where SA and opinion mining comes in. These methods aid in classifying and explaining consumer perceptions towards levels of brands, products, events and services [31]. Notably, SA is

not merely about counting the number of good or bad words but it also considers the sentence structure using a syntactic analysis so as to get the real meaning behind the expressions of the user.

SA refers to the automated sentiment-finding method of analyzing written material of the users, processing the unstructured information, and creating models to derive insights out of it [32]. Most organizations have adopted SA to know how the customer feels about the products or the reaction to the occurrence rather than the most costly and time-consuming method, which is the surveys. The information used in this thesis is collected in Google Play, one of the largest mobile application platform websites, and the individuals tend to leave a review of their experience. Google Play is an unstructured and abundant user-generated content that allows creating millions of reviews across various areas of operation and employs unstructured and informal language and dialects. These qualities make Google Play reviews particularly open to SA, and also pose grave difficulty where data classifiers are concerned and seek to extract substantive information where dealing with text at large scale and noises.

2.4 Natural Language Processing

Natural Language Processing (NLP) is an indispensable component of artificial intelligence that is focused in assisting machine to comprehend, examine and create human language and therefore defines the distinction between human communication and the machine decoding. Under this capability, a ML application in the language processing field can be developed quickly. NLP will be particularly helpful in data science since it can yield valuable data to large amounts of unstructured text produced on a daily basis in medical records, scholarly articles, social media posts, email, and others. NLP initially developed using the systems of linguists who formulated rules to analyze and read the words. Although these rule-based systems were effective for specific, well-defined tasks, their reliance on manually crafted linguistic rules made them difficult to scale, costly to maintain, and poorly adaptable to new domains, languages, or informal and noisy text, which significantly limited their flexibility [33].

NLP has been transformed by the integration of ML and an even more advanced ML, deep learning. These technologies can train NLP systems to learn information, determine trends and make conclusions without human intervention. Such a shift towards more dynamic, data-based

approaches allows the knowledge to be converted into being directly programmed to being extrapolated out of experience. Deep learning models, in particular, those relying on neural networks, including Recurrent Neural Networks (RNNs) and Transformers, can learn to capture the intricacies of the natural language and, therefore, are superior at machine translation, SA, and question-answering capabilities [34].

These inventions have positioned NLP on the edge of the new technologies that are experiencing massive applications, especially in the field of healthcare, where it will be applied in terms of helping to understand clinical stories, extracting patient details, and making decisions [34]. The shift in solving the problems to do with human language is a significant shift in the progression of simple algorithms to more intricate neural formations that has led to the ongoing automation and refinement of the different applications.

Formulation and usage of NLP in computational areas demand the knowledge of the fundamentals of NLP be it syntax, semantics and pragmatics. Syntaxis is related to the rules of the language structure, semantics is related to the meaning of words and phrases under particular circumstances and pragmatics is related to the language use under certain circumstances as a form of seeing the meaning beyond the literal [35]. These have been segmented over NLP techniques like tokenization that breaks the text into small units like words or phrases [36]. Morphological parsing like stemming and lemmatization stretches words down to some basic form, which is needed in interpreting the contextual meanings. Part-of-speech tagging is the determination of grammatical groupings of words and Named Entity Recognition (NER) the designation of valuable data in a text in regard to categories [37].

With the various tools and tactics, it is essential to process and understand human language with the help of these basic aspects. The latter are the primary strategies of NLP that enhance the work of an automated virtual assistant to advanced tools of analysis that could possibly comprehend complex documents and information regarding human behavior and preferences. This complication renders the subject of NLP pertinent in establishing machine interaction and capabilities, which can render them beyond an instrument, though to the extension of human action.

NLP constitutes the core methodological foundation upon which the proposed sentiment and topic analysis framework is built. The study leverages advanced NLP techniques including Arabic-specific preprocessing, deep learning architectures, and transformer-based language models to address the linguistic complexity of Arabic mobile banking reviews. By adapting NLP methods to handle dialectal variation, morphological richness, and noisy user-generated content, this research extends the application of NLP to a domain-specific financial context and demonstrates its practical value in extracting actionable insights for digital banking services in the Arab world.

2.5 Topic Modeling for Arabic Banking Reviews

Topic modeling is an essential element of NLP to discover the underlying latent themes of a large set of textual data, such as digital banking where user-written reviews include a wide variety of unstructured and frequently noisy text. Speaking of the case of the Arabic banking apps, topic modeling proves especially useful in terms of identifying the overarching areas of interest among the users, visualizing the satisfaction factors, and the cross-country differences in the digital service quality task that cannot be exhaustively addressed with the aid of sentiment classification. This layer of analysis supplements SA as it does uncover what and why sentiments arise so that the decision-maker can take more actionable insights. However, the Arabic text topic modeling is highly problematic because of the morphological richness of the language, dialect fragmentation, informal orthography, and the absence of standardized linguistic resources, which is evident in the abundance of the works on the Arabic NLP literature [38, 39, 40] literature. Historically, probabilistic generative models, including Latent Dirichlet Allocation (LDA), are commonly used because they have a mathematical interpretation and are not trained. However, LDA is satisfactory with the structured corpora, but it becomes weaker when subjected to user-generated Arabic texts that often include dialect mixture, spelling inconsistency, implicit sentiment, and short-text documents typical of mobile app reviews. Recent development of LLMs, most notably GPT-4o, has presented a radically new paradigm of topic modeling, based upon semantic understanding, as opposed to the term co-occurrence distributions. GPT-based models are better at contextual reasoning, hence can hypothesize coherent themes even on short or ambiguous text and can create connections across different dialects and linguistic styles. It is becoming clear that topic models based on LLM demonstrate a greater level of performance compared to classical models in coherence, interpretability, and cross-lingual consistency [41]. However, there are still issues of

reproducibility (which may be a hallucination), dependence on fast engineering, and lack of clear probabilistic foundations. Therefore, a hybrid approach to learning traditional statistical models like LDA and GPT-4o-based semantic clustering provides a multi-layered and methodologically sound system to analyze large-scale Arabic banking evaluations. This two-fold strategy allows interpretable assignment of topics as well as high-fidelity semantic understanding, which can help to better comprehend the user behaviors in the multilingual digital banking environment of the MENA region.

2.5.1 Latent Dirichlet Allocation (LDA)

A popular topic modeling algorithm is the Latent Dirichlet Allocation (LDA), which is also used in computational linguistics. LDA is a type of generative probabilistic model that was introduced by Blei et al., [97] documents are characterized as a mixture of latent topics and topics as a mixture of words. The bag-of-words encoding presumes that words are independent, and encoding patterns of semantics are encoded by word co-occurrence frequencies. The interpretability of LDA that is its strongest asset allows the researcher to generate easy to visualize and analyze topic-word distributions, which can be measured with coherence scores like UMass, which measures word co-occurrence based on document frequency within the corpus, and NPMI (Normalized Pointwise Mutual Information), which assesses semantic consistency by capturing how strongly words within a topic are associated across documents. Its performance with English and multilingual corpora is also proven in a number of studies [42, 43] but its capabilities with Arabic texts are limited by the complexity of tokenization, variation in dialect, and low density of documents [98, 99]. This problem is worsened by mobile banking reviews which in most cases contain few words or informal phrases: short texts decrease the performance of LDA since the patterns of co-occurrence are not informative enough. LDA continues to have trouble in situations where subtle semantic relations are in the spotlight. Nevertheless, LDA is a convenient benchmark because it provides clear and replicable distributions of themes and as such it is a fundamental preliminary phase in evaluating thematic patterns in reviews of Arabic banking.

2.5.2 GPT-4o–based Topic Modeling

Large Language Models (LLM) include GPT-4 and GPT-4o as a modern alternative to traditional topic modeling based on probabilistic probing, as they use large-scale contextual knowledge, cross-linguistic semantic mental processes, and extrapolation capabilities during pretraining. GPT-4o proposes topics based on semantic clustering and conceptual abstraction, unlike the use of statistical patterns of word co-occurrence, which allows it to see latent themes in even short, noisy, or dialect-infested text a domain where LDA can perform particularly poorly. Recent publications prove that topic extraction using LLM provides much higher topic coherence, recognizes subtle user behaviors, and multilingual content with greater structure [44, 45]. The ability of GPT-4o to detect sarcasm, idiomatic expressions and dialectal differences, including Gulf, Levantine, or Egyptian Arabic, makes it an effective means of understanding Arabic reviews of mobile banking as linguistic variation is part and parcel of that field. Also, GPT-4o enables dynamic topic modelling with guided prompt engineering enabling researchers to conduct hierarchical topic extraction, sentiment-sensitive topic clustering, and even nation-sensitive thematic segmentation. Nevertheless, there are some shortcomings to the topic-modeling based on the implementation of the LLM: timely stability is a key to reproducibility, and the lack of a clear probabilistic base makes formal evaluation complicated. Moreover, the issues of hallucination, model bias, and interpretability are still present in scholarly studies [46]. Irrespective of these shortcomings, GPT-4o is a transformative innovation in topic modeling, which is currently more accurate and provides a broader context to the review of thematic patterns in Arabic online banking reviews. It allows a hybrid methodological approach that can be interpreted in a balanced way regarding interpretability, semantic precision, and empirical strength when used with LDA.

2.6 Machine Learning

ML is a sub-field of artificial intelligence and data science which focuses on the development of algorithms that are improved by data and experience [47, 48]. When these algorithms are trained on the data sets, they are slowly modified to enhance their ability to make decisions and become more accurate in their prediction [47]. ML is highly embraced in the automation of many tasks and it is classified into different types of learning with each category identified by the type of approach that the algorithm uses to enhance its accuracy.

There are various methods applied to the classification and prediction of opinions, some of them include ML and lexicon based methods; they are very common in opinion mining and prediction [49, 50].

The ML method is oriented at classification, text analysis which are mostly used as assistance to business decision-making and strategic planning. This method involves text pre-processing and firstly makes use of data to learn a model. This model can then make predictions on new datasets that have not been labelled; this has been made possible after the model has been trained. This purpose of ML is divided into two types:

Supervised Learning: This technique involves the use of labeled data to find the patterns and relationships of the training data [52].

Unsupervised Learning: It is implemented to identify trends in unlabeled datasets, and it can provide insights without specified categories [51].

2.6.1 Random Forest

Random Forest classifier is an ensemble model of learning which falls under the category of supraliminal model of learning algorithms [54]. It uses the bagging technique, which has a number of classifiers, where the decision trees are the base learners. The training of each tree in the Random Forest is performed on an arbitrary subset of the training data, sampled randomly, that is, some data points can be repeated. In contrast to certain algorithms that split decisions based on all the features, the Random Forest algorithm uses a random sample of features to build each tree, increasing the variance between trees. This kind of selection does not only reduce correlation between the trees but also reduces the total generalization error. Since trees are trained on varying sets of data, the ensemble is less susceptible to noise and less likely to overfit. Each tree gives a vote during the decision-making process and the majority of the votes get the final output, which is the class, as depicted in Figure 2.1 [53, 54].

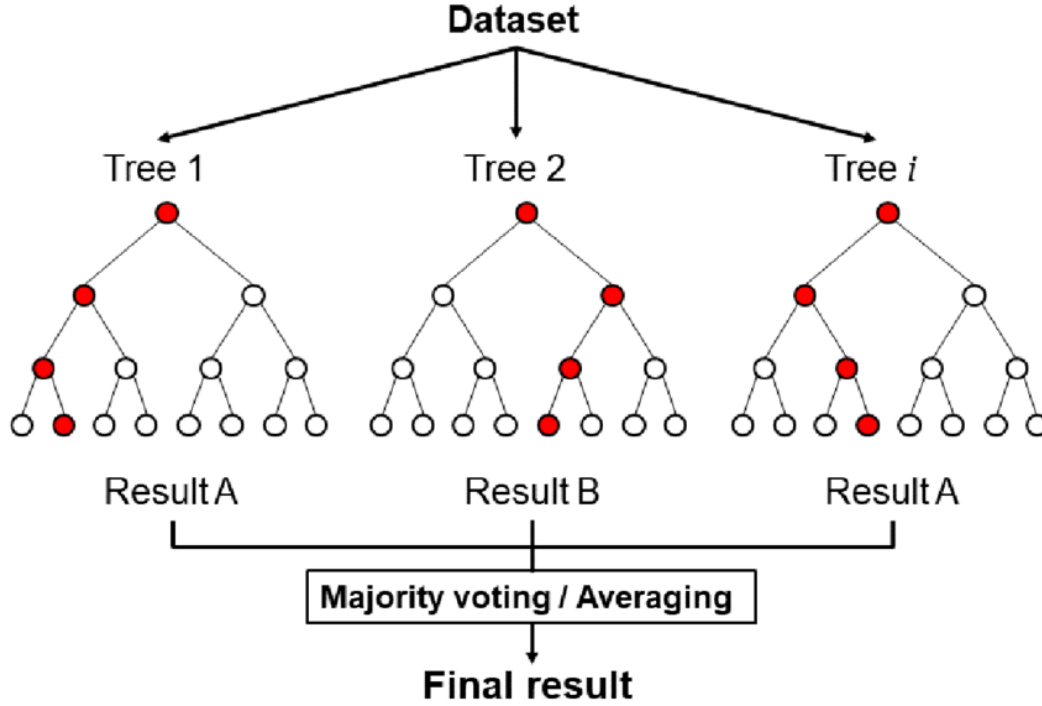


Figure 2.1: Random Forest architecture [55]

2.6.2 Naïve Bayes

Naive Bayes classifier is a kind of supervised learning algorithm that uses the Bayes theorem in its prediction [56]. This theory is used to compute the probability of occurrence of an event given the previously known conditions surrounding the event [57]. Bayes theorem formula is detailed as:

$$P(A | B) = \frac{P(A) \times P(B|A)}{P(B)} \quad (2.1)$$

Here, $P(A|B)$ is the probability of event A given B, $P(A)$ is the prior probability of A, $P(B|A)$ is the likelihood of B given A, and $P(B)$ is the total probability of B. In predictive models, the formula is a product of the previous probability and the probability of occurrence, divided by the products of possible occurrences. Naive Bayes was named so because of the underlying assumption of the algorithm where all the predictors (variables) are independent of one another, a fact that may not always be true in the real world [57].

2.6.3 Logistic Regression

Logistic regression is a method of ML classification, which is especially useful when predicting the likelihood of binary outcomes of a dependent variable [59]. It belongs to the discriminative models of probabilities. These models are based on the assumption that the values of the observed dependent variable are a result of a probability distribution that is given by a function of the features of input. The use of logistic regression is preferred in the modeling of the circumstances in which the response variable is dichotomous. It gives a set of predictions and then tries to apply models to the data that may be determined deterministically and probabilistically [58]. In the event that the training data consists of more than two categories, it is possible to generalize the logistic regression to multinomial logistic regression, which has the ability of predicting more than two class labels.

2.6.4 Support Vector Machines (SVM)

SVMs are considered to be the advanced methods of mining textual information, and they work especially well when dealing with a large amount of textual information [60]. In such cases, SVM and the Naive Bayes classifier are reputed to be efficient and scalable [61]. It is also the distinctive feature of SVM that its regularization is the so-called margin-based separation, i.e. the idea to construct hyperplanes, which would clearly separate points of both groups into two categories. The design of these hyperplanes is symmetric, and they are based on the assumption that the training and test set classes are linearly separable [62].

The process of separating topics into two classes is typically executed using a linear equation:

$$d\mathbf{w} + b = 0 \quad (2.2)$$

Here $\mathbf{w}$ denotes the feature vector as obtained after working on the dataset, $\mathbf{d}$ is the weight vector and b is the bias term provided in the equation.

When data is converted into a form of a vector, it is flexible to any type of data. They operate based on the optimal position of a hyperplane that separates the data space into two different sets that are associated with different classes. This SVM partitioning could be noted in Figure 2.2.

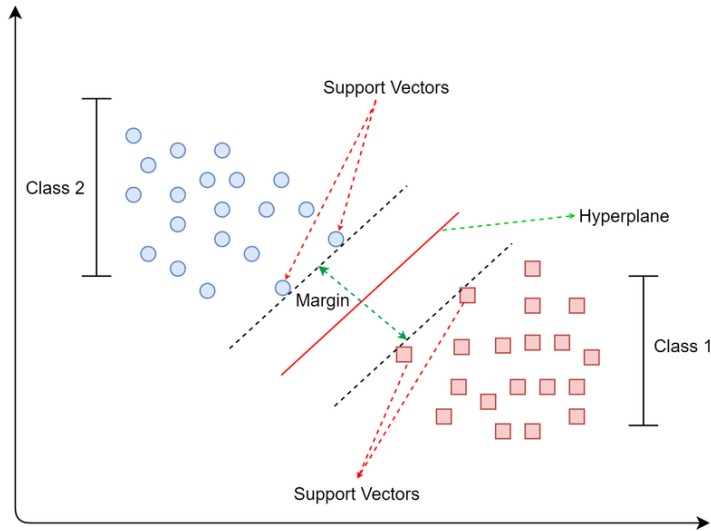


Figure 2.2: Support vector machines architecture [63]

2.7 Deep Learning

Deep Learning (DL) is a subfield of machine learning within artificial intelligence that relies on neural network architectures composed of multiple stacked layers. These layered structures enable models to automatically learn hierarchical representations, where each successive layer captures increasingly abstract and high-level features from the input data [64]. The success of this approach is premised on the fact that it is founded on neural networks which simulate the behavior of the biological neurons in the human brain to accomplish various tasks.

Recurrent Neural Networks (RNNs) are a class of neural network architectures designed to model sequential data by propagating information through recurrent connections. However, standard RNNs often struggle during training due to the vanishing gradient problem, where gradients propagated backward through time gradually diminish, making it difficult to update network weights effectively. To address this limitation, Long Short-Term Memory (LSTM) networks were introduced as a specialized variant of RNNs, incorporating memory cells and gating mechanisms that enable the network to preserve long-term dependencies [65]. The LSTM networks have special categories of structures called gates that control the transmission of information in the network. These gates allow the LSTMs to selectively remember or forget information of previous time steps, and thus are especially useful in tasks such as speech recognition, natural language processing, and time series analysis.

An LSTM network consists of several significant components, i.e., a cell, an input gate, an output gate, and a forget gate. The memory state of the network is included in the cell and it assists in storing important information over a period of time. The input gate decides which information is to be received into the cell, the output gate decides which information to be passed out of the cell to the network output and forget gate decides which information is to be retained or forgotten into the current memory state [66].

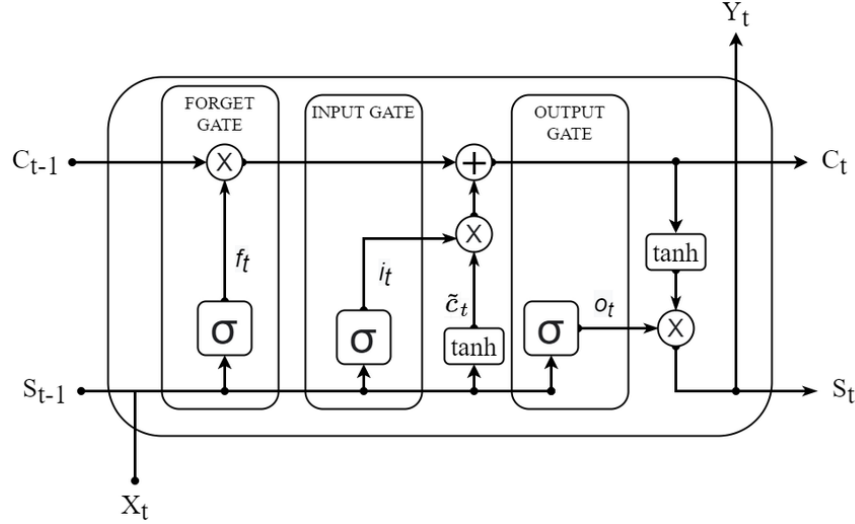


Figure 2.3: Long Short-Term Memory architecture [67]

2.8 Large Language Models (LLMs) for Arabic NLP

Transformer-based Large Language Models (LLMs) have revolutionized the aspect of Natural Language Processing by offering contextual representations that are deep, bidirectional semantic understanding and astonishing generalization to linguistic tasks. The architectures in particular BERT-based ones are founded on the self-attention mechanisms of learning the complex syntactic and semantic dependencies, enabling them to outperform earlier neural and statistical models on SA, topic modeling and dialect identification. Their suitability is especially troublesome in the language that has complicated morphology and a rich inflection, variation in orthograph, and a significant measure of dialect fragmentation, such as the Arabic language. Multilingual models such as mBERT can also be helpful but their performance in Arabic text is low due to comparatively low pretraining exposure, poor tokenization, and cross-language imbalances in training data [28, 29]. They were defeated by Arabic-centric forms of transformer such as AraBERT and MARBERT both of which are better adjusted to the linguistic characteristics and

sociolinguistic diversity of the Arabic text. AraBERT applies large-scale Modern Standard Arabic (MSA) corpora with linguistically aware preprocessing, but MARBERT applies to social media texts and dialectal variation with training on large Twitter corpora. As these models are assembled, one can observe that the direct translation of the specifics of the Arabic language culture and linguistic peculiarities of the language to the direct needs of the model can contribute to the fact that this model may significantly outperform generic multilingual transformers, achieving the state-of-the-art results in sentiment classification, hate speech detection, event extraction, and dialect modeling [11,30, 33]. Despite these strengths, there remain problems, including the intensive cost of pretraining of computing, dialect bias in the subject corpora, and the low interpretability as compared to classical topic models such as LDA. Nevertheless, BERT, AraBERT, and MARBERT are an architectural foundational base of modern Arabic NLP systems that offer powerful architectures capable of processing the informational, user-generated text that is typical of the content of banking applications reviews in the MENA world.

2.8.1 Bidirectional Encoder Representations from Transformers (BERT)

A major breakthrough was also made in the year 2018 when Google researchers introduced the BERT model that was a revolutionary model in the field of Natural Language Processing. The model is constructed on the basis of the Transformer architecture, which was introduced in 2017, and every output is directly tied to all inputs, with a special mechanism, dubbed as Attention, that is used to trace the global dependencies among them [68].

The most important BERT innovation is its training method, which is the pre-training of a deep bidirectional representation of unlabeled text. This approach enables the model to extract meaning of the words before and after the context in every layer on both sides of the word helping it to better understand context and ambiguities in words [69].

BERT architecture is based on the top of a multi-layer two-way Transformer encoder as represented in the original design in figure 2.4. The structure itself is very straightforward since it consists of an encoder-decoder system whereby the encoder receives an input sequence and divides it into continuous representations and the decoder emits output elements one by one [70].

In the original article one notices that the encoder consists of six equal layers, which in turn consist of two major elements as illustrated in Figure 2.4, two separate Multi-Head self-attention

mechanism and a plain feed-forward neural network. These elements are manifested in the decoder with the distinction that it employs the multi-head attention mechanism on the encoder outputs. Besides, a multi-head attention with masks is also integrated to ensure that, at an instance, the predictions can solely be based on the already generated outputs.

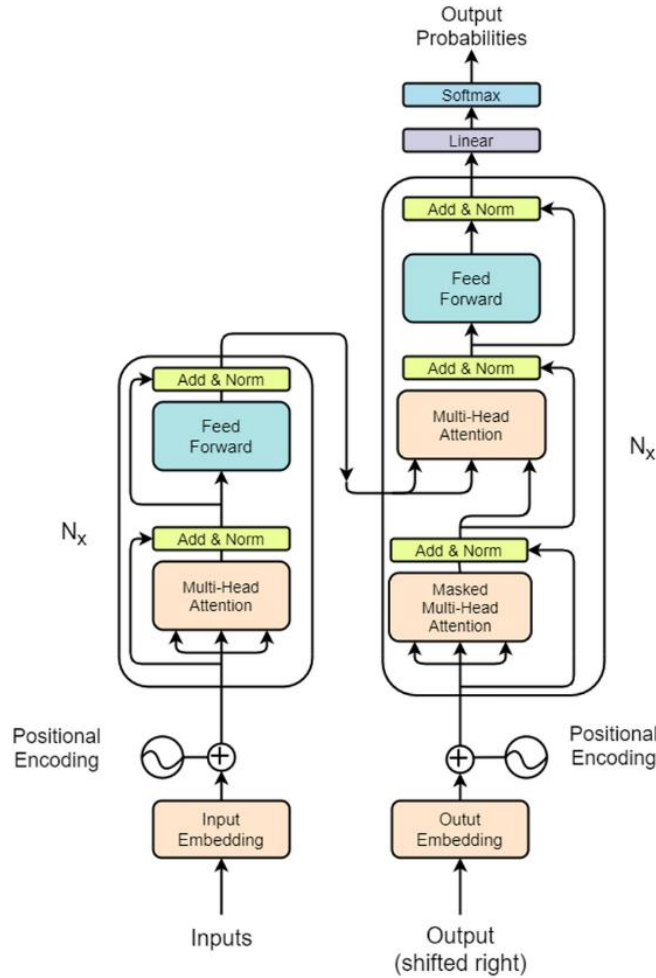


Figure 2.4: Transformer architecture [73]

The attention mechanism is of great importance in this way that model will be able to estimate the relevance or effect between the elements in the sequence. The reason behind this functionality is that there are three types of vectors query, key and value depending on the order of input. By doing so, the relationship process in a sentence is successfully identified using BERT which was highly improved to perceive the situation of a sentence [72].

One of the pros is that the attention mechanism is parallelizable, which has been made possible with the help of the matrix operations that accelerate the calculation of attention. This enables

Multi-Head attention to occur as a result of processing a variety of attention points at once. Figure 2.5 provides a global picture of the architecture of BERT, with the embedding of each input token being computed by a sequence of connected Transformer blocks. This is the repeated structure of the basic BERT model 12 times [71].

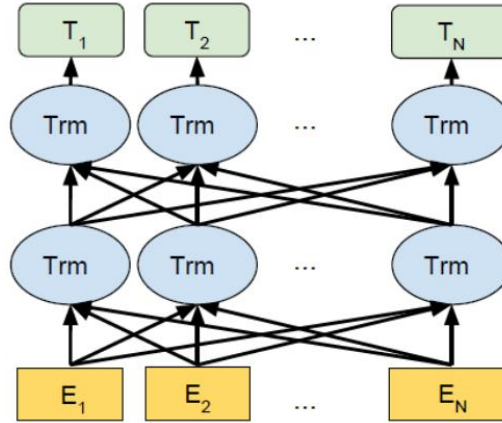


Figure 2.5: BERT architecture [71]

2.8.2 AraBERT

AraBERT has been developed as one of the most influential models of Arabic-only transformers that has been developed to meet the structural and morphological complexity of Arabic language. However, while mBERT was trained on relatively limited amounts of Arabic data, AraBERT was pretrained on large-scale Modern Standard Arabic corpora, including news articles, Wikipedia, and other clean textual sources. The original version of AraBERT was trained on 77 million sentences and 2.7 billion tokens (23 GB of text), which is 17 times larger than the Arabic subset that mBERT is trained on [74]. Later versions still went a step further and made the pretraining corpus 77 GB of text, giving the vocabulary coverage and contextual comprehension a major boost. Architecturally, AraBERT is similar to BERT-base with 12 transformer layers, 768 hidden units, 12 attention heads and a total of about 110 million parameters except that it incorporates the special preprocessing designed to encode Arabic, including diacritic removal, orthographic normalization, and Arabic-aligned tokenization [75] [76]. One further pretraining was done with a 12,000 sentence corpus of dialectal Arabic, which became a customized variant of the model- SalamBERT [77], which improves the model on informal and dialectal text. AraBERT has demonstrated state-of-the-art performance on many Arabic tasks, such as SA, detection of offensive language and text

classification, and is therefore very useful in MSA-intensive tasks like news and formal reviews. Nevertheless, it is less applicable to highly informal, dialect-based social media content as it depends on the MSA data mostly, so it inspired the creation of MARBERT.

2.8.3 MARBERT

MARBERT is a major development in Arabic Large Language Models since the model is availed to address dialect-intensive and social media Arabic, where irregular spelling, bad grammar, emojis, code-switching, and sarcasm are widespread. It was pretrained on a massive dataset of one billion Arabic tweets about 128 GB of text and 15.6 billion tokens which constituted its training set ranked it as the largest Arabic pretraining models to date [78]. MARBERT is the architecturally similar to BERT-base (12 layers, 768 hidden units, about 160 million parameters) only that it does not employ the Next-Sentence Prediction (NSP) goal because tweets are not very long and do not have relevant continuity between sentences.

The distinctiveness of major Arabic dialects, including those of Gulf, Egyptian, Levantine, and Maghrebi, is one of the main strong points of MARBERT. This provides it with the state-of-the-art performance on tasks such as dialect recognition, hate speech detection, sarcasm and SA on informal social media data. Its tokenizer is also user-friendly to online communication, which means that it can better accept hashtags, emojis, long spelling, and informal expressions that are prevalent in user-reviewed app mobile banking services.

MARBERT can be successfully applied to the digital banking analytics where reviewing often fuses dialects and informality by simulating the language patterns of the real users in the Arab world. Although it has its strengths, MARBERT still has the limitation of working on domain-specific financial vocabulary and this can probably be addressed with further domain-adaptive pretraining or even execution of hybrid architectures using large generative models like GPT-4o.

2.9 Related Works

Several studies of SA based on ML have shown a high level of performance on recognizing emotional content using the algorithms of SVM and decision trees. Recent developments in deep learning (DL) and in particular in the use of models such as BERT have further enhanced sentiment detection, as they can recognize contextual and linguistic nuances that traditional models tend to overlook. Simultaneously, the unique challenges of Arabic SA have been the topic of an extensive

research, which dwells upon the problems of complex structure of the language and the lack of linguistic resources. All these methods lead to an understanding of how SA is still developing in new languages based on various computational methods and modeling approaches.

2.9.1 Sentiment Analysis Using Machine Learning

Researchers in [79] used SA to evaluate the user experience of ten mobile banking applications in Pakistan. The experiment was a blend of a valence-conscious sentiment dictionary and ML methods. The Multinomial Naive Bayes, Logistic Regression and Support Vector Machine were tested with an ensemble model to compare the performance. They were also extracted in the research by the use of the Top2Vec model to get thematic factors that affect the usability of these applications. The ensemble classifier gave the best results with an F1-score of 90%. The thematic analysis revealed 346 positive themes, including reliability and easy navigation, and 441 negative themes including frequent issues, like technical failures, high performance, and crashes of apps.

In [80], researchers suggested a method of the comparative evaluation of the ML algorithms used in the context of SA of Twitter data by classifying the claims in three categories, namely, positive, negative, and neutral. TF-IDF and multinomial NB, SVM, and LR utilized different methods. The results showed that the LR worked best particularly when used with n-gram and bigram models with an accuracy of 86.23 being higher than what was obtained with SVM and multinomial NB (85.69% and 83.54% respectively).

Recent research has focused on identifying the most effective feature extraction techniques for sentiment analysis, emphasizing the critical role of representation methods in model performance [42]. In this work, several widely used approaches—namely Bag-of-Words (BoW), Word2Vec, n-grams, TF-IDF, Hashing Vectorizer (HV), and Global Vectors for Word Representation (GloVe)—were systematically evaluated. These feature representations were applied to two benchmark datasets: Twitter reviews related to U.S. airlines and Amazon musical instrument reviews. A Random Forest classifier was used to assess performance across methods. The experimental results showed that TF-IDF outperformed the other techniques, achieving accuracy scores of 99% on the Amazon dataset and 96% on the Twitter dataset. These findings highlight the decisive impact of feature extraction choices on the effectiveness of sentiment analysis models.

In [81], researchers suggested that SA be handled by using the methods of ML techniques to limit the number of reviews that are overwhelming in e-commerce. They employed several classifiers, such as SVM, NB, Decision Tree, RF, LR and Multi-layer perceptron (MLP) to do supervised learning on the reviews of the products of the Amazon Electronics. Also, they contrasted three preprocessing techniques are TF-IDF, Bag of Words, and Word2Vec, to model review text. The most effective technique was the MLP classifier which was used together with Bag of Words preprocessing, which achieved an accuracy of 92%.

In [82], researchers suggested to apply supervised ML algorithms in SA of customer and employee feedback to increase profitability of the organization and comprehend needs in a better manner. They compared many algorithms including linear regression, KNN, SVM, RF, bagging and gradient boosting using data, which comprised 802 trustworthy employee responses on an original population of 884. The sentiment was assessed with the help of AFINN, VADER, and TextBlob, and the aspects of leadership that were considered were strength, perspective, and recognition. The research concluded that RF algorithm was the most accurate with a value of 0.711 which was very high compared to KNN whose accuracy was 0.515.

In [83], researchers suggested a paradigm of multilingual SA of both Arabic and English, making use of two datasets (ASTD, AJGT) and translations. Some of the preprocessing methods they used include n-gram tokenization, stop words removal, removal of punctuations among others. Four ML classifiers, namely LR, RF, NB and SVM, were put to test. The research has looked at how preprocessing affects the accuracy of the two languages. The findings demonstrated that SVM worked best on ASTD dataset with an accuracy of 70 percent whereas NB had the highest accuracy at 87 percent on the AJGT dataset. There was no sign of performance difference in the translated datasets but a few features performed slightly better in Arabic.

In [43], the researchers suggested comparing the sentiment classifiers being SVM, NB, LR, and RF to enhance the SA techniques. They investigated the effects of different Natural Language Processing (NLP) pipelines and how sentiment prediction can be improved with the help of lexical semantic knowledge of WordNet on three datasets with total 50 000 movie reviews. The experiment showed that various NLP strategies and semantic improvements have a significant

impact on performance with the highest accuracy having lemmatization and knowledge-based n-gram features with the highest score of 90.43% by the SVM.

Based on the above related works of SA using ML techniques, there are limitations that can be identified. To begin with, most of the studies have employed more traditional models like Naive Bayes, SVM, and Logistic Regression; these models are to an extent naive relative to deep learning models as the models fail to understand the intrinsic language semantics. Moreover, the efficiency of models can depend on the preprocessing and filtering of the input in some techniques that are unique to a specific language, e.g. TF-IDF or Bag-of-Words. This may complicate cross-dataset and even cross-language translation. The model performs poorly on multilingual datasets, and we observe heterogeneity in the results in the presence of language-specific aspects, as well as, preprocessing approaches. Furthermore, when it comes to the accuracy values reported in numerous subsequent studies, they seem to be high frequently but do not necessarily have the full applicability to the real-world due to overfitting on particular data sets and/or limited scope of tested data. Lastly, a more universal problem is that none of the studies mentioned above comments on the scalability of these models, or how their models are suitable to real-time analysis which is a requirement of many real-world applications.

2.9.2 Sentiment Analysis with Deep Learning

Researchers in [84], researchers evaluated the performance of Google’s FNet model, which replaces the attention mechanism used in BERT with a Fourier transform to improve computational efficiency in NLP tasks. The study fine-tuned FNet on a Kaggle hotel review dataset and compared its results with BERT, LSTM, and SVM classifiers. The outcomes showed that FNet reduced training time by about 20% and cut memory consumption by nearly 60%, while still reaching a test accuracy of 80.27%—approximately 97.85% of BERT’s performance under the same conditions.

In a similar line of research, the authors of [85] presented a hybrid ensemble DL sentiment analysis architecture based on RoBERTa paired with LSTM, BiLSTM, and GRU. RoBERTa is herein charged with the responsibility of developing contextual embeddings whereas the recurrent models obtain long-term language dependencies. Ensemble averaging, majority voting and addition of GloVe pre-trained vectors were also used to make the system further enhanced to solve the problem

of class imbalance. The hybrid model produced the best results than the current methods with 94.9% accuracy on IMDb, 91.77% on the Twitter US Airline dataset, 89.81% on Sentiment140.

The authors presented the SLCABG model in [86] as a SA model that incorporates a sentiment lexicon, a Convolutional Neural Network (CNN), and attention-based BiGRU. Sentiment lexicon serves to enhance the feature representation and CNN and BiGRU layers are used to capture contextual information as well as sentimental cues. There is then an attention layer which classifies the most relevant features. In 100, 000 reviews of products in the Chinese e-commerce site, dangdang.com, the SLCABG model yielded an accuracy of 93.5, indicating a high level of accuracy in SA of Chinese texts.

Researchers have introduced a DL model based on Bidirectional LSTM-RNN model with attention mechanisms to study the sentiment of COVID-19 related Twitter content in [87]. The model was trained on 179,108 tweets of a Kaggle dataset that has four emotional categories (sadness, joy, fear, and anger). The use of attention layers enhanced the performance greatly in that it increased the accuracy by 20 percent, enhanced the precision by 10-12 percent and the recall was enhanced by 12-13 percent over the earlier methods. The paper also found that half of the tweets were positive, forty-five percent of the tweets were neutral, and five percent were negative, which proved the effectiveness and practicability of the proposed model in classifying the sentiments related to the pandemic.

In [88], researchers proposed a model of consumer review summarization that utilizes NLP methodology along with an LSTM-based system to process the text in several formats: social media, e-commerce platforms, and news websites. The suggested system is a three-step hybrid design: the first one stages the removal of the noisy or redundant data with the help of preprocessing, the second stage identifies the general review features as well as review aspect features, and the third step involves the utilization of LSTM classifier to identify sentiment. The model had an average precision of 94.46, a recall of 91.63, and an F1-score of 92.81, meaning that when tested on three separate datasets, it has given companies valuable information about the preferences and behavior of the consumers.

The work presented in [89] introduced B-MLCNN, a BERT-enhanced Multi-Layered Convolutional Neural Network designed for document-level sentiment classification. This model has been applied in areas such as e-commerce and cloud computing to categorize sentiment into positive, negative, or neutral classes for improved decision support. The architecture uses a pre-trained BERT model to generate word-level feature vectors, which are then processed through multiple convolutional layers with varying kernel sizes to extract deeper semantic patterns. These feature maps are fused and passed to a Softmax layer to perform final sentiment prediction. Experimental results on IMDB and Amazon review datasets showed that B-MLCNN achieved accuracy levels near 95%, demonstrating strong potential for real-world applications.

In [90], the authors described the real-time SA method using Twitter that is suitable to process large amounts of data in a short period of time. Examples of common preprocessing include stemming, tokenization and stopword and irrelevant word removal, which form the workflow. The optimized data that is currently processed is stored with the Hadoop Distributed File System (HDFS) and MapReduce to remove redundancy. Extraction of features is done on aspects like emoticons and other indicators of expressiveness that have been widely used in twitter. To facilitate classification, the paper uses a Modified DL Neural Network (DLMNN), which proved to be better than the conventional models such as ANN, SVM, K-Means and DCNN. The suggested model demonstrated good assessment scores with a precision of 95.78 percent, recall of 95.84%, a 95.87% F-score, and 91.65% overall accuracy.

The restrictions that is evident in the studies that have been done on advanced SA models indicate that there are a number of challenges that still exist. Despite the significant improvements to accuracy and efficiency of such approaches as FNet and BERT-based hybrids, their practical applicability is still limited due to high computational costs and training data that cannot be easily extended to new text domains or to other languages. The biggest challenge is still the requirement of large well-annotated datasets, when such models are used on smaller, noisier or ill-structured data then the performance can decrease vastly.

Furthermore, although recent models are able to score well on benchmark datasets, whether they can support real world content, particularly in text which can contain sarcasm, ambiguity or context-specific sentiment is an open question. Even with the development of model architectures,

such as attention mechanisms and ensemble strategies, various systems still experience challenges such as overfitting and poor transferability to different sentiment domains and applications.

2.9.3 Sentiment Analysis for Arabic Language

The authors in [91] developed a weakly supervised system, called q8SentiLabeler, to address the difficulties of manually labelling sentiment data in the Kuwaiti dialect. The large dataset of 16.6k social media posts constructed allowed them to classify the sentiments without focusing heavily on human labeling and just restricted bias to language variation and temporal changes. Both the conventional ML models and the sophisticated DL models were used to assess the reliability of the dataset, and q8SentiLabeler demonstrated a high level of pairwise agreement at 93% and a Cohen kappa of 0.87. The ARBERT model achieved accuracy of 89 when applied on this dataset, indicating the viability of this weakly supervised model.

In [92], researchers have suggested a system that will improve the accuracy and efficiency of the Arabic SA, especially when it comes to customer reviews of coffee products on Twitter. They employed a hybrid feature extraction method to overcome the linguistic complexity with the Arabic morphology which combined TF-IDF and Minimum Redundancy Maximum Relevance (MRMR) on 10,646 tweets. It used four supervised algorithm classifiers namely k-nearest neighbor (KNN), SVM, decision tree, and random forest with ensemble voting algorithms. The hybrid system was found to have significant improvements, recording a high accuracy of more than 95.95 with hard voting and approximately 94.51 with soft voting.

The study in [93] compared quantum computing models with classical ML for Arabic document classification. On a large dataset of 213,465 tweets, quantum approaches slightly surpassed conventional ML in sentiment tasks and completed processing roughly one minute faster. On a smaller dataset of 44,000 tweets, both approaches executed more quickly, with classical ML producing marginally better accuracy and evaluation metrics. The results illustrate that quantum computing is a viable and competitive option for large-scale SA, particularly in terms of processing efficiency.

In [94], researchers recommended the assessment of the performance of single and combined ML algorithms in foul language detection and cyberbullying of Arabic text. Three Arabic datasets were

used with three ML classifiers and three ensemble models applied to them, two of which are publicly available and the third one is specifically designed. It was discovered that the ensemble techniques, especially the voting classifier, were superior to the single learners. The voting ensemble had a better performance of 71.1%, 76.7% and 98.5% on the datasets, than that of the best single learner of 65.1%, 76.2% and 98%. Hyperparameters to the Arabic cyberbullying dataset were further optimized.

In [95], researchers suggested a comparative study of three ML classifiers namely LR, K-Nearest Neighbors (KNN), and Decision Tree (DT) in analyzing sentiment about Arabic tweets utilizing four datasets. Results showed that LR was the most accurate especially when working with large data sets where accuracy was up to 93%. KNN and DT, conversely, experienced a noticeable decline in performance with the largest dataset and the accuracy dropped to 74 percent. Also, LR and KNN had higher effectiveness, compared to DT, on smaller datasets.

In [96], researchers put forward a multi-criteria procedure to analyze and rank the classifiers in Arabic SA, with the help of ML algorithms. Evaluation was done by conducting reviews of Saudi Arabian products to compare between five classifiers. The findings showed that the DL and SVM classifiers were better as the accuracy of the former was 85.25%, the precision was 85.30%, the recall was 88.41%, the F-measure was 86.81%, the classification error was 14.75% and the AUC was 93%. SVM was very close behind, showing good performance in all measures, hence performing better than KNNs, DT and NB classifiers.

In the articles reviewed in the research on Arabic SA, there are a number of outstanding issues that are based on the linguistic and contextual complexity of the language. A significant challenge is based on the rich morphology of Arabic and the existence of a vast amount of dialects that are used in addition to Modern Standard Arabic. This therefore indicates that models which has excelled in one dialect fail to generalize to the others causing a break in their capacity to contain the entire lingual diversities existing among Arabic speaking nations. The other limitation is that it makes use of datasets that are either too restricted or have been captured in one platform that they cannot reflect the overall attributes of Arabic text. Together with the intensive use of manual annotation, the research process is also limited by the lack of well-labeled data, especially those related to dialects that are not well studied.

Also, although certain novel methods have shown good improvements, most of the sophisticated classification methods still have high computational complexity, which cannot be easily implemented in practical systems. These problems suggest that more solid, scalable, as well as versatile models of SA are required that can adequately address the entire continuum of Arabic linguistic diversity, as well as contextual specifics.

3. Methodology

3.1 Introduction

This chapter proposes a multi-step approach to the creation of a consistent framework for SA and topic extraction of Arabic banking app reviews. The framework will involve the collection of data on Google Play, Arabic-centric preprocessing and stratified sampling to maintain even distribution of sentiments. To promote the quality of labels, a hybrid labeling approach is adopted to incorporate the use of manual review, rating-based tags, and GPT-4o support, whereas data imbalance issues are resolved using class weighting, SMOTE, oversampling, and undersampling. The data processing algorithm uses TF-IDF to extract features and tests several models, such as SVM, Random Forest, LSTM, and AraBERT. It ends with statistical and semantic topic modeling to expose the concealed themes to create a full pipeline that is sensitive to the linguistic richness of the Arabic language and the diversity of user-generated materials.

3.2 Proposed Model for Arabic Sentiment analysis of Banking Application Reviews

A large number of the currently used methods address the issues of Arabic SA one at a time, and this can be particularly prone to preprocessing, annotation, and model evaluation inconsistency. In order to effectively deal with these issues in a holistic manner, this research proposes a systematic and extensive pipeline incorporating high quality preprocessing, hybrid labeling, balanced sampling, and multi-model evaluation. The complete architecture to be used to analyze sentiment of the Arabic reviews of mobile banking apps is shown in Figure 3.1.

The processing model starts with the user feedback gathered from Google Play and then consistent processing stage specific to the linguistic peculiarities of Arabic. This involves normalization of the text, elimination of URLs, punctuation marks, special characters, diacritics, mixed letters, unnecessary spacing, duplicates, and blank lines to maintain the quality of data. A stratified sampling method is then selected to give a balanced group of reviews to be annotated.

Labeling is performed through a combination of expert manual annotation, GPT-4o-assisted contextual classification, and automated sentiment inference derived from the numerical user ratings available on Google Play, where star ratings are used as an initial polarity indicator to

support sentiment assignment. In order to reduce the class imbalance, three balancing measures are tested separately, namely class weighting, SMOTE, and random oversampling.

After cleaning and labelling, the data is converted to numerical features with the help of TF-IDF and divided into training and testing data, and K-fold cross-validation is used to provide fair and reliable evaluation. The extraction of topics is processed with the LDA and the GPT-based semantics modeling to demonstrate the important themes in the user feedback.

To be classified, the study focuses on three classes of models, such as traditional ML algorithms (SVM, Naive Bayes, Random Forest, Logistic Regression), LSTM, and state-of-the-art Arabic language models (AraBERT, MarBERT). Lastly, the performance of the models is strictly tested on the basis of generally accepted measures to determine the most efficient methods of predicting sentiments in Arabic banking application reviews.

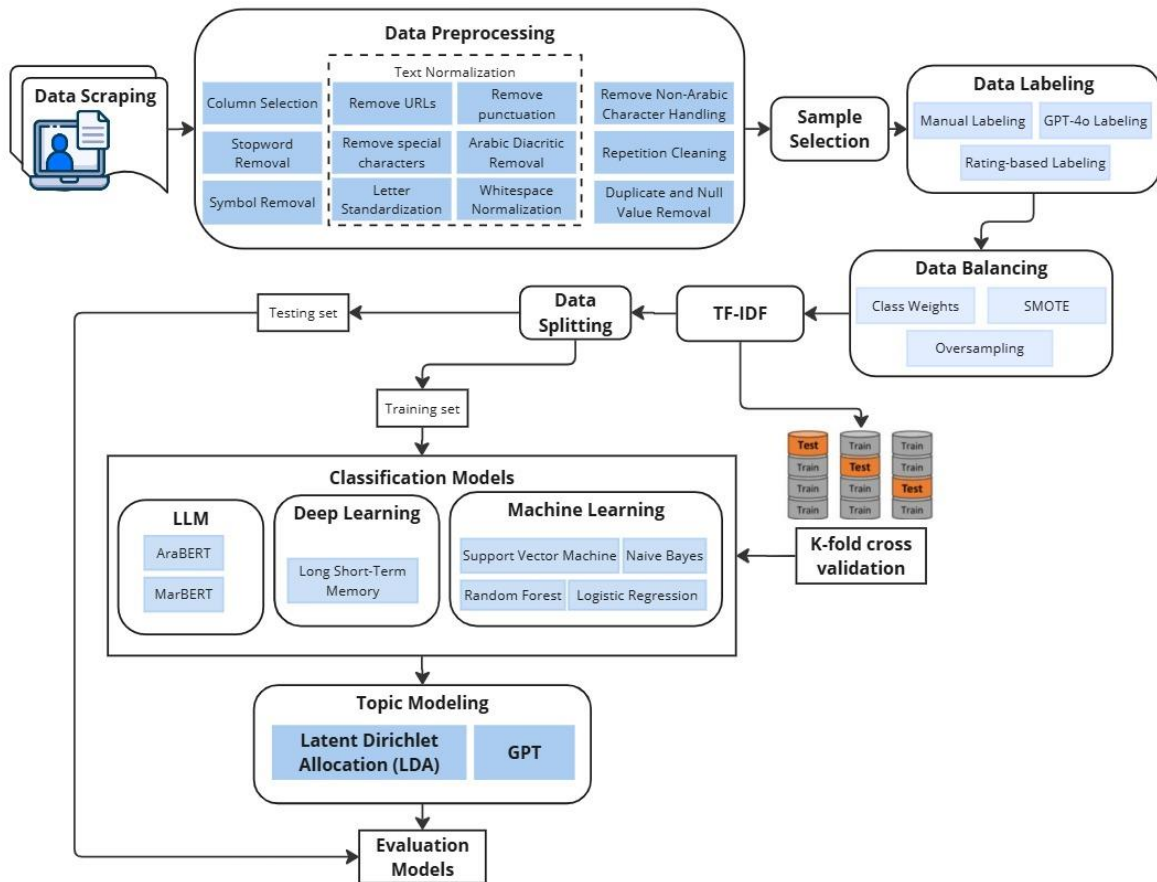


Figure 3.1: The Proposed Framework for Arabic Sentiment Analysis of Banking Application Reviews

3.2.1 Data Scraping

The collection of data was one of the key points of this study and was based on large-scale collection of user reviews on Google Play about mobile banking applications associated with the top 100 Arab banks chosen by The Banker in 2024 [100]. The procedure generated over 361 000 reviews, and, originally, the reviews were not restricted to any language. To ensure the data is consistent with the research aim, the reviews were subsequently filtered to include only Arabic contents such that analysis was done on Arabic-speaking user reviews only.

Most of the reviews were based in Saudi Arabia and had more than 270, 000 entries. Jordan (23,499), Egypt (14,781), and the United Arab Emirates (14,781) ranked higher and there were very few representatives of such countries as Lebanon and Tunisia (165 and 64 reviews respectively). Such differences are indicative of the actual situation in terms of mobile banking use and digital adoption in the MENA region. Instead of artificially equalizing the number of users in the region, the original proportions were kept to retain the original user behavior and permit an insightful exploration of the sentiment variation among various dialects and demographic groups. The maintenance of this natural distribution is also helpful in the creation of models that are capable of managing linguistic diversity of Arabic dialects which is needed to construct SA tools that are generalizable across communities and banking users.

3.2.2 Data Preprocessing

Preprocessing text is an essential step towards any NLP process, and in the case of Arabic, a language that can be characterized by confusing morphology, spelling variations and dialects, this step is particularly important. In this study, a well-planned preprocessing pipeline was developed to modify the raw user reviews into uniform and useful text to be analyzed. This was not just to eliminate noises and redundancy, but also to save on expressions that have a lot of emotion and dialectal indication that contribute to proper classification of sentiment. Every stage of the pipeline addressed a particular linguistic issue in the dataset.

A sample of the raw reviews that were scraped on Google Play is shown in Figure 3.2, and it is important to note that the following fields were considered throughout its processing: the source, the review_description, the rating, and the country_code. All these characteristics create contextual information required to understand the sentiment and detect language-specific features. This example also demonstrates the plurality of user-generated content among various countries

and banks, as writing styles, expressions of sentiment and the use of dialect are very diverse among Arabic mobile banking feedback.

	source	review_description	rating	country_code
0	Ahli United Bank	ممتازة	4	Bahrain
1	Ahli United Bank	٢٠٢٤ / ٨ / ١٨...م يطلع لي تنبيه آتة غير آمن!! مدر	1	Bahrain
2	Ahli United Bank	ممتازة	5	Bahrain
3	Ahli United Bank	تمام	5	Bahrain
4	Ahli United Bank	رائع	5	Bahrain
...	...	...	...	...
361940	BankMuscat	كيف طريقة الاستخدام	5	Oman
361941	BankMuscat	شكرا	3	Oman
361942	BankMuscat	اصلي.ولا.مزيف الفلوس بيروحن في النهاية	5	Oman
361943	BankMuscat	يجب التأكد هل البرنامج رسمي من البنك أو مزيف	1	Oman
361944	BankMuscat	ما ندري اذا برنامج اصلي ولا مزيف	3	Oman

361945 rows × 4 columns

Figure 3.2: Dataset collected from Google Play

Multi-layer text normalization process was implemented to deal with the variations of spelling and structural noise of Arabic user reviews (Table 3.1). The initial phase was aimed at eliminating such factors that have no meaning bearing, including URLs, excessive punctuations, random symbols, and non-informative special characters such as decorative symbols, repeated emojis, formatting artifacts, and isolated characters that do not contribute to the linguistic or contextual meaning of the text. The diacritics were also dropped to cut the number of vocabularies into pieces and widely misinterpreted letters like the alif in various forms, ya, ta marbuta were standardized to one unified form. The repetition and irregular spacing was further fixed and made clean and regular in order to have a cleaner and more consistent text structure.

Stopword removal was subsequently performed, with special attention to preserving negation words like “ما”, “مو”, “لا”, and “ليس”, since these terms directly affect sentiment polarity. Unlike generic filtering approaches, this method safeguarded against incorrect classification that could arise if negation markers were removed. Simultaneously, a non-Arabic character filter was applied,

which removes the needless foreign characters and retains valuable English words, numerical data, equipment and version names commonly used in banking application reviews.

More optimization was carried out by removing the redundancy of visual effects and the overindulgence of emotive expression of users. Normalization of repetition of letters and word elongation were normalized into a standard form without meaning or emotional impact. Lastly, empty records and duplicate records were eliminated to ensure data quality, which led to a significant decrease in the number of repeat reviews by more than 210,000 and the removal of over 18,000 empty records. The final corpus of reviews after cleaning consists of 147,338 distinct reviews, and both raw and processed versions are available to ensure transparency and facilitate future comparative analysis.

Table 3.1: Examples of Arabic Text Before and After Each Preprocessing Step

Preprocessing Step	Before	After
1. URL Removal	التطبيق يعلق كل دقيقة! جرب بنفسك http://testbank.com	التطبيق يعلق كل دقيقة! جرب بنفسك
2. Punctuation & Special Character Cleaning	خدمة البنك سيئة مزعجة !!??	خدمة البنك سيئة مزعجة
3. Arabic Diacritic Removal	تطبيق بسيط وسهل الإستخدام جداً	تطبيق بسيط وسهل الاستخدام جدا
4. Letter Standardization	أرجووك حلوا المشكلة	ارجووك حلوا المشكلة
5. Whitespace Normalization	خدمة العملاء ممتازة جداً	خدمة العملاء ممتازة جدا
6. Stopword Removal (Negation Preserved)	ما في تحديث جديد و ليس سريع	ما في تحديث جديد و ليس سريع
7. Non-Arabic Character Handling	Bug بس فيه Face ID التطبيق يدعم	Face ID التطبيق يدعم Bug بس فيه
8. Symbol Removal	!!!خدمة* البنك?? ممتازة	خدمة البنك ممتازة

9. Repetition Cleaning	!!رائع جداااa
------------------------	---

Figure 3.3 shows the result of the entire preprocessing process of the data, which shows the final structured set of data after all cleaning operations. The sequential normalization and filtering processes led to a harmonious and uniform text, which is a strong basis of precise SA and valuable topic extraction.

	source	review_description	rating	country_code	Cleaned_Review	word_count
0	Ahli United Bank	م ٢٠٢٤ / ٨ / ١٨...مطلع لي تنبيه أنه غير آمن!! مدر	1	Bahrain	١٨ ... ٢٠٢٤م مطلع تنبيه انه غير امن مدري خلل صرنا	18
1	Ahli United Bank	جيده جدا	5	Bahrain	جيده جدا	2
2	Ahli United Bank	الضيق لا يعمل علي الجهاز	1	Bahrain	الضيق لا يعمل علي الجهاز	5
3	Ahli United Bank	لايعمل بصورة جيده	1	Bahrain	لايعمل بصورة جيده	3
4	Ahli United Bank	...تم تثبيت التطبيق أكثر من مره ولا يريد يشتغل عل	1	Bahrain	...تم تثبيت التطبيق اكثر من مره ولا يريد يشتغل عل	16
...	...	...	...	...	...	...
149083	BankMuscat	جميل .. ولكن كيف تسجل فيه ؟؟؟	5	Oman	جميل ولكن كيف تسجل	4
149084	BankMuscat	كيف طريقة الاستخدام	5	Oman	كيف طريقة الاستخدام	3
149085	BankMuscat	اصلي..ولامزيف الفلوس بيروحن في النهاية	5	Oman	اصلي ولا مزيف الفلوس بيروحن النهاية	6
149086	BankMuscat	يجب التأكد هل البرنامج رسمي من البنك أو مزيف	1	Oman	يجب التأكد هل البرنامج رسمي من البنك او مزيف	9
149087	BankMuscat	ما ندري اذا برنامج اصلي ولا مزيف	3	Oman	ما ندري اذا برنامج اصلي ولا مزيف	7

149088 rows × 6 columns

Figure 3.3: Dataset After Preprocessing

3.2.3 Sample Selection

In order to make the SA fair and meaningful, a stratified sampling approach was used because there was an overwhelming imbalance within the raw data. Figure 3.4 shows that most of the reviews were concentrated around the 1-star and 5-star ratings, which are typically associated with negative and positive sentiment labels, respectively. This distribution reflects a common behavioral pattern in online review platforms, where users are more likely to leave feedback when they experience extreme satisfaction or strong dissatisfaction, while moderately satisfied users (e.g., 2–4 stars) are less inclined to submit reviews. This phenomenon explains both the polarization of ratings and the implied relationship between rating values and sentiment polarity.

A stratified random sampling method was adopted to minimize bias and assist in creating stable classification models. This approach maintained the percentage representation of all the rating levels by taking 5,000 reviews in each of the five rating levels and came up with a balanced sub set of 25,000 reviews.

This made selection balanced to avoid the effect of sentiment dominance by extreme reviews, enabled the models to learn equally across categories, and provided a reasonable comparative point during evaluation. The sampling procedure was also done using a fixed random seed (random_state = 42) to ensure that the sampling process is reproducible in future experiments. Any further annotation and modeling was done on this balanced dataset which will guarantee fair training and more valid performance evaluation.

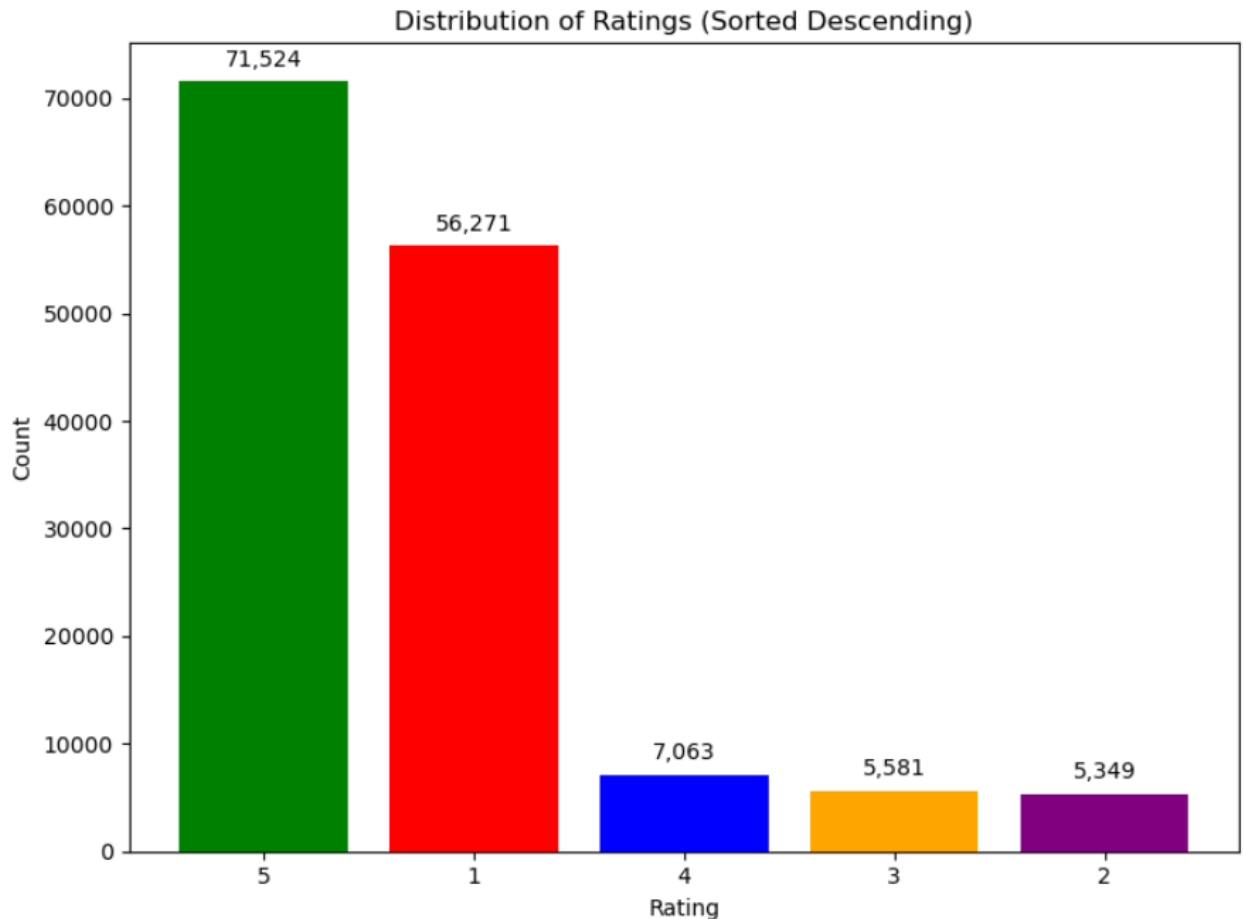


Figure 3.4: Rating Distribution of Mobile Banking Application Reviews Before Sampling

3.2.4 Data Labeling

There was a multi-source labelling procedure after a sample selection, which was created to tackle the fine sentiment articulations present in Arabic user-generated content. Three antithetical labeling strategies were applied to the balanced subset in each of which the benefits were provided, and a good multi-perspective sentiment ground truth was examined.

3.2.4.1 Manual Labeling

The initial and most authoritative technique was manual annotation and it was conducted by the author, a native Arabic speaker with professional experience working in the banking sector and strong familiarity with banking terminology as well as various Arabic dialects. This anthropomorphic labelling model has been used as the gold standard of the research because it relies on human judgment in interpreting some subtle signals of language, some indicators of aversion that are not explicit, sarcasm, idiomatic expressions, and dialect-specific variations which are often poorly recognized by automated systems. Manual labeling is associated with high quality and contextually correct sentiment assignment, as well as with a critical standard of automated labeling technology evaluation.

It is important to acknowledge that the manual annotation was performed by a single annotator, which represents a limitation in terms of inter-annotator reliability, particularly given the dialectal diversity present in the dataset. While the annotator possessed native Arabic proficiency and professional banking experience, individual annotation inherently introduces subjectivity, especially when interpreting sarcasm, dialectal expressions, and culturally specific language. Accordingly, the manual labels should be regarded as reference labels rather than absolute ground truth. Future work could strengthen annotation quality by employing multiple annotators and measuring inter-annotator agreement using Cohen's Kappa or Krippendorff's Alpha

3.2.4.2 GPT-4o Labeling

The second labeling method utilized the use of OpenAI GPT-4o, a large language model (LLM) that is capable of proficiency in both the Modern Standard Arabic and regional dialects. The GPT-4o was used as an easy and efficient alternative to manual annotation in auto-annotation of their reviews. Its contextual reasoning can reason out subtle sentiments such as sarcasm, negation, indirect complaint, mixed opinions which often occur in Arabic reviews of mobile applications. The high level of semantic knowledge of the model enables it to be useful in the benchmarking of

the performance of the traditional ML models in sentiment classification. It is important to note that GPT-4o serves two distinct and non-overlapping roles in this research: first, as an auxiliary annotation tool whose labels are compared against manual annotations to assess inter-method agreement, and second, as a semantic engine for topic modeling. GPT-4o is not employed as a sentiment classifier in the model evaluation pipeline. All sentiment classification models (SVM, Logistic Regression, Naive Bayes, Random Forest, LSTM, AraBERT, and MARBERT) are trained and evaluated exclusively on the manually annotated reference labels. Consequently, there is no model circularity, as GPT-4o does not influence the training data or the evaluation outcomes of the sentiment classifiers.

3.2.4.3 Rating-Based Labeling

The third process employed a rule-based process by getting the sentiment out of the numerical rating of the respective reviews. The ratings of 4 and 5 were considered as Positive (+1), rating 3 was considered as Neutral (0) and the rest of them, 1 and 2, were considered Negative (-1). In spite of the linguistic and contextual sensitivity of manual and GPT-based annotation, this approach offers rapid and scalable baseline consistent with much of the previous SA practice. It is simple and thus quick to initially label, but also has some chances of inconsistencies, especially where the textual information does not wholly contextualize the numerical rating.

All three labeling strategies, manual, GPT-4o, and rating-based categorized the chosen reviews into Positive, Neutral, and Negative in the same three categories of sentiments. The differences in the number of classes in these approaches represent the natural disparities in the language interpretation, contextual inferences and philosophy of annotation.

3.2.5 Data Balancing

There was a need to define the target sentiment categories before applying any balancing technique to ensure model stability, linguistic interpretability, and robust classification performance. The dataset originally consisted of three sentiment classes; however, since the neutral category was not aligned with the objectives of this study, the number of categories was reduced to two: positive and negative reviews. The neutral class was considered non-essential because the primary aim of the research is to distinguish customer satisfaction from dissatisfaction in order to generate actionable insights for banking service improvement, whereas neutral reviews typically contain ambiguous, factual, or low-emotion statements that do not clearly express evaluative attitudes.

The distinction between positive, neutral, and negative reviews was determined through the multi-source annotation framework adopted in this study, which included manual expert labeling, GPT-4o-assisted classification, and rating-based inference, with manual annotation serving as the gold standard to ensure consistent and linguistically grounded sentiment differentiation.

In practical Arabic banking feedback, neutral reviews often consist of vague statements, ambivalent expressions, or brief comments that do not convey a definite emotional orientation. Such cases introduce noise, increase annotation uncertainty, and reduce the discriminative clarity required for effective model learning. Furthermore, the variability observed in the labeling of neutral reviews across annotation methods—particularly between manual and GPT-4o labeling—reduced their suitability for reliable multi-class training. Binary sentiment classification also aligns more closely with the operational goals of banking analytics systems, which primarily focus on identifying customer satisfaction versus dissatisfaction to support decision-making and service enhancement. Accordingly, transforming the dataset into a binary format (1 = Positive, 0 = Negative) provided a clearer training signal, reduced ambiguity, and improved interpretability of the results.

After removing neutral entries from the stratified balanced subset (25,000 reviews obtained by selecting 5,000 reviews from each of the five rating levels), the remaining manually validated dataset used for binary modeling consisted of 13,914 negative and 9,664 positive samples (total = 23,578), with the difference attributable to the exclusion of reviews labeled as neutral during the annotation consolidation phase. This distribution reflects a moderate level of class imbalance that required systematic correction.

To address the remaining imbalance between classes, four balancing methodologies were evaluated individually to assess their impact on model performance: class weighting, SMOTE, random oversampling, and random undersampling. Each technique represents a structured strategy for compensating class inequality and was applied independently to ensure a fair and controlled comparison within the Arabic sentiment classification framework.

3.2.5.1 Class Weights

Class weighting represents a model-level correction that adjusts the importance of each class during training while leaving the underlying data unchanged. Under the balanced class-weighting technique, the majority class (negative) and the minority class (positive) were assigned weights of

0.85 and 1.22, respectively. This adjustment helped normalize the influence of both sentiment categories and resulted in an approximately balanced effective contribution of about 11,789 samples per class during model optimization (Figure 3.5). It is a computationally inexpensive and strategically efficient balancing approach, as class weighting does not require modifying the dataset itself and is particularly effective for linear models such as Logistic Regression and SVM.

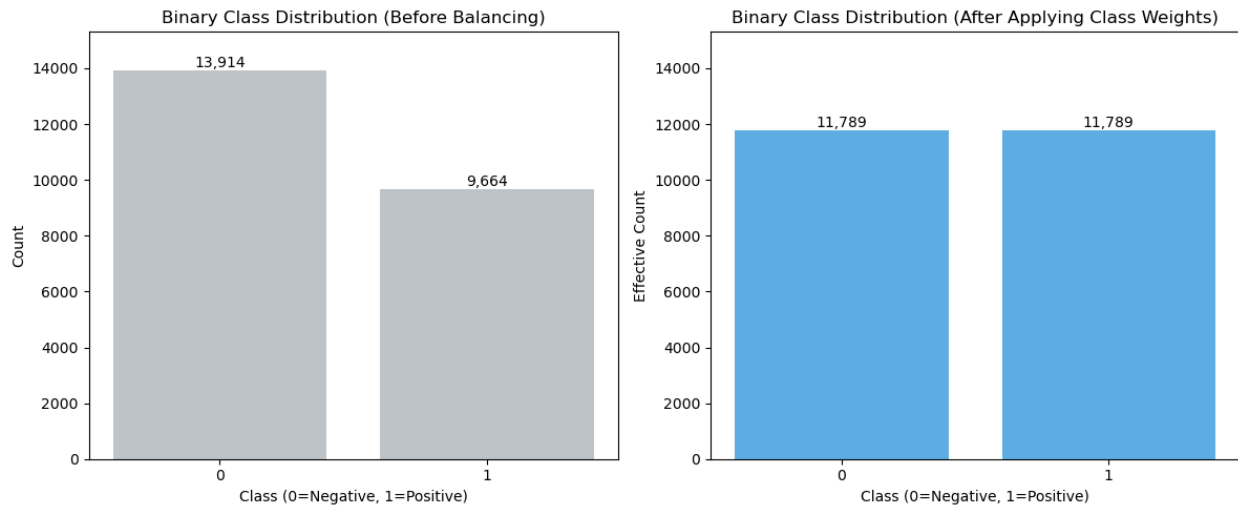


Figure 3.5: Class Distribution Before and After Applying Class Weights

3.2.5.2 SMOTE (Synthetic Minority Oversampling Technique)

Synthetic examples were created in order to fill the gap of positive samples in the training data by the means of the SMOTE algorithm. This was applied only to the training set to ensure that the test set remained unchanged and that the evaluation was not biased. In the beginning, the training data included 11,131 negative reviews and 7,731 positive reviews. Following the application of SMOTE, the number of samples in the two categories was equal at 11,131 each, so that a total of 22,262 reviews is now included in the training, as indicated in Figure 3.6. SMOTE makes new samples by interpolating the feature values in the minority class. This can avoid overfitting and enhances the generalization capacity of the model to capture a significant benefit in the case of linguistic diversity and structural complexity of Arabic user reviews.

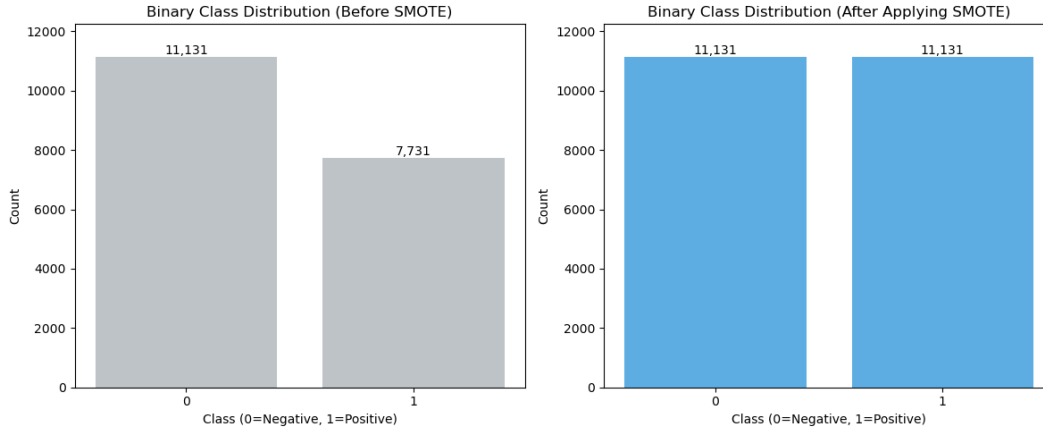


Figure 3.6: Class Distribution Before and After Applying SMOTE

3.2.5.3 Random Oversampling

Random oversampling is also another technique of counterbalancing imbalance in that the examples belonging to the minor classes are duplicated. There is no artificial noise of the approach and no distortion of the original patterns of the linguistic structure of the positive reviews. This was then oversampled and the training set consisted of 22, 262, one to one, (11, 131/each) of the two classes (Figure 3.7). This method is easy, very quick and efficient in low density high-dimensional representations like TF-IDF matrices, but can result in overfitting in certain models because of repeated cases.

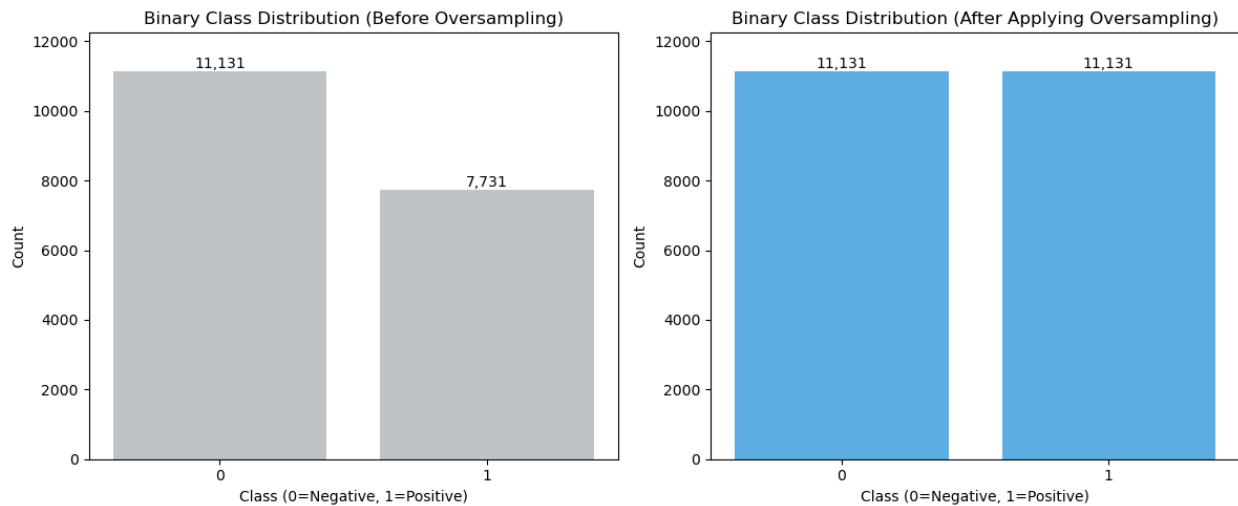


Figure 3.7: Class Distribution Before and After Applying Oversampling

3.2.5.4 Random Undersampling

Random undersampling was done to remove the majority (negative) class randomly until it was equal to the minority class. The initial training set of 11,131 negative and 7,731 positive reviews was then narrowed down to 15,462 samples with an ideal distribution of 7,731 negative and 7,731 positive reviews, whereas the test sample remained the same (Figure 3.8). Though computationally inexpensive and ensuring complete class balance, undersampling can cause some possibly informative examples to be lost and the total diversity of the dataset to be poorer--a factor to take into account when the morphological diversity experienced by Arabic text has to be dealt with.

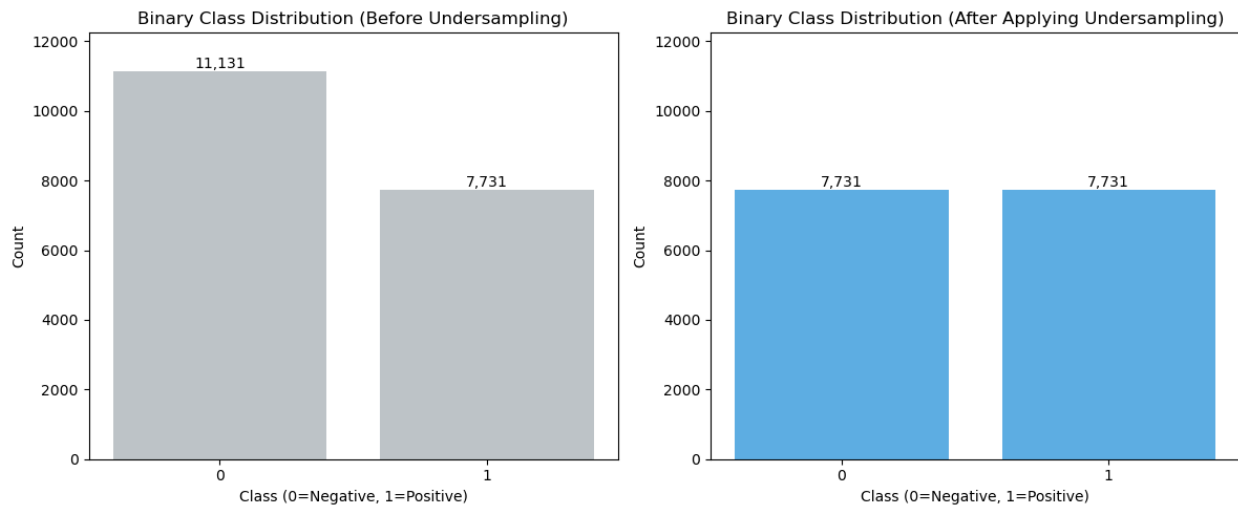


Figure 3.8: Class Distribution Before and After Applying Undersampling

3.2.6 Term Frequency–Inverse Document Frequency (TF–IDF)

TF-IDF is a method that is commonly employed to transform texts into numerical features when performing NLP tasks, including SA and document classification. It gives a weight to each word which is the ratio of how many times in a particular document word occur in comparison with how many times in the entire data they occur. This is calculated using two terms which are Term Frequency (TF) which is the frequency of word usage in a document and Inverse Document Frequency (IDF) which minimizes the impact of too common words which do not provide much discriminatory information. TF-IDF score of a term t within a document d is defined as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \quad (3.1)$$

The term frequency $\text{TF}(t, d)$ measures how often a word appears within a document:

$$\text{TF}(t, d) = \frac{\text{Number of occurrences of term } t \text{ in document } d}{\text{Total number of terms in document } d} \quad (3.2)$$

To reduce the weight of words appearing frequently across many documents—such as common stopwords—the inverse document frequency $\text{IDF}(t)$ is computed as:

$$\text{IDF}(t) = \log \left(\frac{N}{1 + \text{DF}(t)} \right) \quad (3.3)$$

where N is the total number of documents and $\text{DF}(t)$ represents the number of documents containing the term t .

The combination of TF and IDF gives the representation an accent of words, which are significant in a review, yet few in a corpus. In this thesis, TFIDF was specifically good at dealing with noisy, informal, and dialect-rich Arabic Google Play reviews. It enriches sentiment-sensitive keywords and reduces repetitive or non-informative words and therefore it is a powerful representation base of ML models and an indispensable component of subsequent analyses like topic extraction and feature interpretation.

3.2.7 Machine Learning Models for Sentiment Classification

The proposed SA plan implies the application of several classic ML models to develop a set of strong baseline models and evaluate the conventional algorithms performance in case of handling large scale and dialect intensive Arabic Google Play reviews. Naive Bayes, SVM, Logistical Regression and random forest are models representing different learning paradigms that can be used to provide a full assessment of the non-linear and linear text classification capabilities. All models are trained on TF-IDF vectorized features and include ways of dealing with class imbalance, in particular by involving class-weight adjustments that penalize misclassified individuals of minority classes equally.

3.2.7.1 Multinomial Naive Bayes

The Multinomial Naive Bayes classifier is a probabilistic classifier founded on Bayes' theorem, with a simplifying assumption that the features are conditionally independent with respect to the class label. This assumption means that, given a specific class (e.g., positive or negative), the occurrence of one word is treated as independent from the occurrence of another word. In other words, dependencies among features are not modeled directly; instead, each feature contributes individually to the probability of the class. The independence is therefore defined relative to the class label, not in absolute terms among the features themselves.

Irrespective of this strong assumption, Naive Bayes remains a competitive classification tool in text classification, particularly in scenarios involving sparse and high-dimensional textual data. This is because text representations such as TF-IDF generate large feature spaces where most values are zero, and in such conditions the independence approximation often performs well in practice despite being theoretically restrictive.

The model estimates the posterior probability of a sentiment class by approximating the distribution of term frequencies within each class. The posterior probability refers to the probability that a given review belongs to a particular sentiment class (e.g., positive or negative) after observing its words. It is computed by combining the prior probability of the class with the likelihood of the observed terms under that class, according to Bayes' theorem. This probabilistic formulation enables efficient training and rapid inference.

Although Naive Bayes does not inherently support class-weight adjustments in the same manner as some discriminative models, its computational efficiency, interpretability, and stable baseline performance make it an appropriate benchmark for comparing more complex machine learning and deep learning models.

3.2.7.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most effective conventional models for sentiment analysis, particularly when combined with TF-IDF features. In this research, a LinearSVC classifier is employed with a linear kernel, which is well suited for large and high-dimensional feature spaces typical of textual data. The objective of SVM is to determine the optimal separating hyperplane that maximizes the margin between positive and negative sentiment classes, thereby improving generalization performance. Maximizing the margin increases robustness to noise and

reduces the risk of overfitting, as the decision boundary is positioned at the greatest possible distance from the nearest training samples (support vectors) of both classes.

The training process incorporates class weights to address dataset imbalance, ensuring that misclassification of minority-class instances incurs a higher penalty. The combination of margin maximization and class-weight adjustment results in a discriminative and robust classifier capable of handling the noise and variability inherent in user-generated Arabic text.

3.2.7.3 Logistic Regression

Another popular discriminative text classification model that is more often used to provide interpretability and good results in binary sentiment classification is the Logistic Regression. The model estimates the probability that a review belongs to a specific sentiment class using the sigmoid function applied to a linear combination of the input features. Logistic Regression is used in the proposed framework with a higher maximum number of iterations to achieve convergence on the high dimensional TF-IDF inputs. Logistic Regression is frequently used as a competitive background because it is much more efficiently computed, gives probabilistic results, and can capture linear decision boundaries in text.

3.2.7.4 Random Forest

The Random Forest classifier implements a non-linear learning approach by constructing an ensemble of decision trees, where each tree is trained on a bootstrap sample of the dataset and a random subset of features. This strategy reduces variance and mitigates overfitting, making the ensemble more stable and robust than a single decision tree. The proposed system employs a Random Forest model with 200 estimators, a maximum tree depth of 20, and defined minimum samples required for node splitting and leaf formation. These hyperparameters were selected to balance model complexity and generalization performance while maintaining computational efficiency. Specifically, limiting the tree depth controls excessive growth that may lead to overfitting, while using multiple estimators enhances robustness through aggregation of diverse decision boundaries.

Class-weight adjustments are incorporated to address class imbalance, ensuring that both sentiment categories are adequately represented during training. These weights are applied during

the model fitting stage on the training set, where they influence the impurity calculation at each node split, thereby assigning greater importance to minority-class samples without altering the original dataset.

Random Forest also provides measures of feature importance, offering interpretability by identifying the most influential textual features contributing to sentiment prediction. In addition, its ensemble structure enables the model to capture complex and non-linear relationships within the data, which may not be effectively modeled by linear classifiers.

3.2.8 LSTM for Sentiment Classification

Classic RNNs are capable of processing sequential text; however, they struggle with long-term dependencies due to the vanishing gradient problem, which limits their effectiveness in modeling natural language. LSTM networks address this limitation through memory cells and gated mechanisms (input, forget, and output gates) that regulate the flow of information, enabling relevant contextual signals to be retained over longer sequences. This capability is particularly important for Arabic sentiment analysis, where sentiment polarity may depend on delayed negation markers, idiomatic expressions, or contextual cues that appear later in the sentence.

The proposed LSTM-based deep learning architecture is designed to capture temporal dependencies within Arabic sentences as well as their latent semantic representations. The model begins with an Embedding layer that transforms each token into a 128-dimensional dense vector representation. This dense embedding space enables the model to learn contextual similarity between words and capture semantic relationships across dialectal and morphological variations common in Arabic text.

The embedding output is passed to a stacked LSTM architecture. The first LSTM layer consists of 128 units with `return_sequences=True`, allowing it to output the full sequence of hidden states rather than a single vector. This configuration preserves sequential information and enables the second LSTM layer to model higher-level temporal abstractions across the entire sentence. The second LSTM layer, composed of 64 units, processes this sequence and extracts more compact temporal representations.

To improve generalization and reduce overfitting, a Dropout layer with a rate of 0.1 is applied after each LSTM layer and after the dense layer. Dropout randomly deactivates a fraction of neurons during training, preventing co-adaptation of units and enhancing the model's robustness on unseen data.

The architecture then includes a fully connected Dense layer with 64 neurons and ReLU activation, which transforms the extracted temporal features into a discriminative representation suitable for classification. Finally, the output layer consists of a single neuron with a sigmoid activation function, producing a probability score between 0 and 1 for binary sentiment classification.

This architectural design was selected to balance representational capacity and computational efficiency, making it suitable for moderately sized Arabic review datasets. The LSTM model serves as a strong deep learning baseline for sequential modeling and allows comparison with transformer-based Arabic language models, which rely on attention mechanisms rather than recurrent structures. Rather than claiming superiority, this study positions LSTM as an interpretable and computationally tractable benchmark within the experimental framework.

Table 3.2: LSTM Model Architecture and Parameters

Layer	Description / Configuration
Embedding Layer	Vocabulary size = <i>max_words</i> ; Embedding vector size = 128; Input length fixed to <i>max_len</i>
First LSTM Layer	128 memory units; return_sequences=True; Uses tanh for hidden state and sigmoid for recurrent gate activation
Dropout Layer	Random dropout probability: 0.1
Second LSTM Layer	64 memory units; return_sequences=False; Hidden activation: tanh; Recurrent activation: sigmoid
Dropout Layer	Dropout rate set to 0.1
Fully Connected (Hidden) Layer	64 neurons; Activation function: ReLU

Dropout Layer	Dropout applied with a rate of 0.1
Final Output Layer	1 neuron; Sigmoid activation for binary sentiment classification

3.2.9 AraBERT for Sentiment Classification

AraBERT is a transformer-based language model specifically developed for Arabic Natural Language Processing tasks. It is built upon the original BERT architecture, which consists of stacked bidirectional Transformer encoder layers utilizing self-attention mechanisms. Through bidirectional contextual modeling, AraBERT captures both left and right context simultaneously, allowing it to learn rich semantic and syntactic representations of Arabic text.

AraBERT is pretrained exclusively on large-scale Arabic corpora, including Modern Standard Arabic (MSA), news articles, and publicly available social media text. This language-specific pretraining enables AraBERT to better model Arabic linguistic characteristics such as complex morphology, rich inflectional patterns, orthographic variation, and dialectal diversity. As a result, it can learn more fine-grained linguistic representations that are particularly beneficial for downstream tasks such as sentiment analysis, text classification, and named entity recognition.

In addition, AraBERT employs an Arabic-specific WordPiece tokenizer that is adapted to handle Arabic morphological structures. This tokenizer segments words into subword units that preserve meaningful morphemes, enabling improved handling of affixes, clitics, and morphologically complex word forms compared to generic tokenizers designed for multilingual settings. Consequently, AraBERT provides more linguistically appropriate representations for Arabic text processing tasks.

The model used in this thesis builds upon the linguistic strengths of AraBERT v0.2 (bert-base-arabertv02) and is fine-tuned for binary sentiment classification of Google Play application reviews. Rather than modifying the core architecture of AraBERT, the pretrained model is fine-tuned by adapting its final classification layer to the specific downstream task of binary sentiment prediction.

The first stage involves specialized tokenization, where all reviews are processed using the pretrained WordPiece tokenizer associated with AraBERT. Truncation and padding are applied

with a maximum sequence length of 256 tokens to ensure uniform input dimensions for the transformer encoder. The resulting inputs are encoded as `input_ids` and `attention_mask`, and are organized using Hugging Face Dataset objects to enable efficient batching and GPU-based processing.

The classification framework directly leverages the `AutoModelForSequenceClassification` implementation provided by the Hugging Face Transformers library, eliminating the need for a custom classifier module. AraBERT functions as a contextual encoder that generates deep bidirectional representations of the input text. The final classification head consists of a linear layer producing two logits corresponding to the positive and negative sentiment classes.

Model optimization is performed using the standard `CrossEntropyLoss` function applied to the output logits during the forward pass, ensuring effective parameter updates for binary sentiment classification without introducing additional architectural modifications.

`TrainingArguments` provided by HuggingFace are used to set the necessary fine-tuning process with the prudent choice of hyperparameters. Training is performed with a batch size of 16 and evaluation is performed with a warmup phase of 200 steps to stabilize gradient updates in the early stages of training. To prevent overfitting, a weight decay of 0.01 is applied and to monitor the performance of the evaluation loss, the validation is conducted after each epoch and the best model is automatically stored in case of the evaluation loss. The given training plan will ensure the stable convergence and the optimal optimization of AraBERT to the linguistic peculiarities of the Arabic app-review reviews. The suggested AraBERT structure presents a powerful and language-sensitive approach in sentiment classification with large-scale, dialect-varied Arabic Google Play data via transformer-based contextual model, imbalance-sensitive learning, and domain-specific fine-tuning.

3.2.10 MARBERT for Sentiment Classification

MARBERT is a transformer-based language model specifically designed to address the linguistic diversity of Arabic, particularly as used on social media platforms. Developed by researchers at the University of British Columbia, it builds upon the BERT architecture and is pretrained on large-scale Twitter and user-generated datasets, enabling it to effectively process dialectal variations, informal spelling, and noisy online text. It employs a WordPiece tokenizer capable of handling

complex Arabic morphology, making MARBERT particularly suitable for sentiment analysis of informal content such as tweets, application reviews, and, in this study, Arabic Google Play reviews.

In the proposed framework, MARBERT begins with text preprocessing and tokenization using its pretrained tokenizer. Each review is converted into input representations consisting of `input_ids` and an `attention_mask`, with uniform padding and truncation applied to a maximum sequence length of 256 tokens to ensure consistent input dimensions. These encoded representations are then organized into Hugging Face Dataset objects to enable efficient batching, memory management, and parallel processing during GPU-based training.

Class weights are computed to address training data imbalance, allowing the model to penalize misclassification of minority-class samples more strongly. MARBERT is fine-tuned using `AutoModelForSequenceClassification`, with a weighted `CrossEntropyLoss` applied during training to maintain sensitivity to class distribution. The MARBERT encoder generates contextualized embeddings, which are passed to a classification head that outputs logits for the two sentiment classes: positive and negative.

The training is performed using Hugging Face `TrainingArguments`, which is a configuration class in the Transformers library that defines key training parameters such as learning rate, batch size, number of epochs, evaluation strategy, logging behavior, and optimization settings, with a strict configuration designed to ensure stable convergence and optimal adaptation to the specific characteristics of Google Play review data. The model is trained for four epochs with a batch size of 16 for training and 32 for evaluation, and includes 200 warmup steps to stabilize gradient updates during the initial phase of optimization. A weight decay of 0.01 is applied as a regularization technique to reduce overfitting by penalizing excessively large parameter values. Model evaluation is performed at the end of each epoch, and automatic checkpointing based on validation loss is employed to retain the best-performing model version during training.

The specified training strategy will assist in ensuring that MARBERT will be capable of accommodating the trend of sentiment that will be present within the app-based user reviews without losing the ability to read dialectal peculiarities and informal language usage signs. The proposed MARBERT architecture offers a powerful and linguistically conscious approach to

Arabic sentiment classification with satisfactory options of coping with the diverse and disordered nature of Arabic digital data in the real world.

3.2.11 Topic Modeling

In order to support the sentiment classification part of the given framework, topic modeling methods are used to reveal the hidden themes that inform the experiences of users when visiting Arabic banking applications. The collected data consist of very brief and noisy texts due to the short length and dialect-rich nature of the reviews. Therefore, this study combines a conventional probabilistic modeling approach based on Latent Dirichlet Allocation (LDA) with an advanced topic extraction method using a Large Language Model (LLM) implemented through GPT. LDA estimates corpus level thematic distributions in a structured way whereas GPT estimates more coherent semantically aware topic discovery, especially of informal and context-dependent Arabic text. These methods combined will provide a holistic view of the concerns of the users, allowing a more detailed look at the aspects of the motivations of the positive and negative attitudes.

3.2.11.1 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic algorithm that is popular in natural language processing as an attempt to discover latent theme representations in large bodies of text. It presumes that documents consist of a compound of a topic, each subject being represented in its turn by a distribution of words. This asset renders LDA specifically appropriate to learn the underlying themes that influence user experience in unstructured and high-volume data like Arabic banking application reviews.

LDA has been selected in this thesis due to a number of reasons. First, LDA provides an interpretable topic–word structure, which is essential for deriving practical insights from customer feedback. Second, LDA performs effectively on noisy and short texts compared to more complex neural topic models and transformer-based approaches, making it a suitable choice for dialect-rich Arabic reviews where linguistic variation is significant. Third, LDA allows positive and negative sentiments to be modeled separately, enabling the extraction of sentiment-specific themes that provide detailed insights into aspects that users appreciate or find problematic.

The evaluation of LDA within this framework follows a rigorous pipeline consisting of multiple preprocessing and modeling steps to ensure reproducibility and thematic consistency. The process

begins with the importation of the required libraries and the setting of fixed random seeds to guarantee deterministic and reproducible results.

The labeled data is later loaded, cleaned and filtered to only keep the preprocessed Arabic reviews in binary sentiment categories. The special preprocessing step is operated, involving tokenization, deletion of stopwords, which have exceptions (negation cues), and producing bigrams representation to reflect multi-word formations, which are typical in Arabic user reviews. Different dictionaries and bag-of-words corpora are built concerning positive and negative reviews and the LDA learns specific topics relevant to each sentiment category. In order to identify the best number of topics, coherence scores (c_v metric) are calculated over a predetermined range making sure that the models selected are maximizing semantic coherence. After finding the best amount of topic numbers (15 in positive reviews and 14 in negative reviews), multiple passes of final LDA models are trained on adaptive hyperparameters (alpha and eta) to produce models with improved topic stability and reduced noise. The topics obtained are then extracted, manually labelled in Arabic and English and organized into distribution tables about the frequency and prevalence of those topics in each sentiment. Lastly, various visualization modes such as coherence plot, topic distribution plot, word cloud grid as well as PyLDAVis interactive map are incorporated to offer an exhaustive as well as understandable thematic examination. Taken together, this LDA-based topic modeling method provides a strong and interpretable tool of making sense of the repetitive themes within the Arabic banking application reviews and supplementing the sentiment classification models with more qualitative information.

3.2.11.2 Generative Pre-trained Transformers (GPT)

Generative Pre-trained Transformers (GPT) are a novel large language model that builds on deep transformer models, large-scale pretraining on large multilingual datasets, and semantically coherent and context-dependent text outputs due to the use of rich attention mechanisms. GPT models are good at making inferences about minor semantic relationships, receiving implicit themes, and simulating human-natural language descriptions of latent concepts. This usability of GPT especially in the analysis of Arabic user-generated content, whose linguistic complexity, such as dialectal variation, inconsistent spelling, code switching, or common usage of colloquial language, creates a serious challenge to traditional topic modeling algorithms.

The choice of GPT in this thesis is prompted by its unmatched capability to identify interpretable and context-sensitive topics in short reviews and generate consistent topic names and descriptions and key words, which are much closer to human judgment. Contextual reasoning of GPT allows it to find themes that are not explicitly mentioned but can be guessed based on sentiment, usage trends, or user intent, which is more insightful and more useful in the analysis of how customers perceive banking applications.

The proposed framework incorporates GPT by using a well-structured batch-processing pipeline that can be scaled effectively and reliably with over ten thousand Arabic reviews on Google Play. It starts with the review segmentation into positive and negative groups, which allows sentiment-specific topic identification. The subsets are further split into reasonable chunks (80 reviews at a time), which will enable a to-the-point interaction with the GPT-4o API. In each batch, a structured prompt is built, which includes all the review texts and instructions to instruct the model to extract one to three important topics in a review. The prompt provides consistency by requesting GPT to produce a Markdown-formatted table that includes: the Review ID, an English topic name, a brief English description, and a list of Arabic keywords directly extracted on the basis of the review. This makes the extracted topics be rooted with the original text and that they represent significant semantic units.

To mitigate the limitations of API and achieve robustness, the pipeline has automatic retries, autosaving of batch results, and is systematically regenerating results in case of interruptions. After processing all the batches, the extracted topics are aggregated into a single set of data, duplicate eliminated by combining the Review_ID and topic name filter and further cleaning undertaken to maintain thematic accuracy. Topics created using GPT on positive and negative reviews are then combined to create a large topic repository, which includes user praise, concerns, and recurring issues.

In order to increase usability, the English topic labels are batch-translated to formal Arabic with GPT-4o-mini, which generates academically appropriate terms to be used in reporting. Lastly, analytical methods are enforced to calculate the frequency of topics and extract the top fifteen of each sentiment category, which is followed by visualization that accentuates the prevalent topics. This pipeline enables GPT to act as both an efficient topic extractor and a powerful linguistic interpreter, providing some degree of semantic richness and thematic unity that model-based

approaches are unable to provide, thereby contributing substantially to the quality of insights gained by reviewing the reviews of Arabic banking apps.

3.3 Conclusion

This chapter has summarized the whole methodological model that was used in construction of a robust and linguistically based system of sentiment classification and topic modeling of Arabic banking application online reviews. A series of data acquisition, preprocessing, sampling, multi-source labeling, class balancing, feature engineering, and model implementation helped the study provide a solid basis of analyzing large-scale, dialect-rich, and structurally diverse user-generated output. The combination of the classic ML methods, the DL structures, and the sophisticated Arabic language models will guarantee a full-scale analysis of the sentiment detection methods, and the presence of probabilistic and semantic-based topic modeling will allow gaining a better insight into the issues raised by users and their thematic patterns. Every methodological choice in the treatment of orthographic variation to the choice of balanced training strategy was constructed in order to maximize the validity, transparency, and repeatability of further analysis. With the now created methodological framework of the research, the next chapter will present the experimental findings and performance appraisals of the models proposed in the present work, providing an elaborate analysis of their efficacy in sentiment representation and gainful extraction of meaningful issues through Arabic banking reviews.

4. Results and Discussion

4.1 Introduction

This chapter introduces and discusses the empirical evidence based on the user reviews of the mobile banking applications that are conducted in Arabic. The findings are structured in a way that they give an all-round assessment of the suggested sentiment classification pipeline after which there is the semantic meaning of user feedback with the help of topic modeling. First, the experimental design is described, including the computational environment, libraries, and method of evaluation to support methodological rigor and reproducibility. The chapter then assesses the reproducibility of annotation decisions between various labeling methods, creating a good ground truth to use in downstream. Then, ML, DL, and LLMs are compared based on their classification performance using conventional performance metrics. Lastly, the topic modeling findings are analyzed using LDA and GPT based thematic interpretations in order to identify the underlying patterns and drivers of user experience in the reviews collected. A combination of the analyses gives a combined view of sentiment behavior, usability issues, as well as functional expectations in Arabic mobile banking ecosystems.

4.2 Experimental Setup

4.2.1 Implementation Environment

Each of the experiments performed in the thesis was run on Python in Google Colab Pro environment, which offers scalable cloud computing functionality that is perfectly adapted to ML processes.

Google Colab had several benefits such as GPU/TPU acceleration, longer memory allocation, and direct support of Python scientific libraries as well to provide fast execution of high-dimensional vectorization, model training and evaluation tasks.

Moreover, Colab enabled the establishment of stability and reproducibility of versions, through enabling fixed library versions to be installed, so that all experimental outcomes are reproducible under the same conditions. The Python implementation offered some flexibility in text preprocessing, model development, and visualization and was compatible with popular SA tools and ML frameworks.

4.2.2 Libraries and Tools

Several Python libraries were also used to feature different elements of the proposed SA framework. Data preprocessing, structured data handling, and numerical operations were performed using Pandas and NumPy, while advanced Arabic text preprocessing tasks—such as stopwords removal and linguistic normalization—were facilitated using the Natural Language Toolkit (NLTK).

The backbone of ML was scikit-learn with support the implementation of SVM, Logistic Regression, Naive Bayes, Random Forest, and assessment tools like train-test splitting, cross-validation, TFIDF vectorization, and performance metrics. To deal with class imbalance processes, Imbalanced-learn (imblearn) gave strong implementations of SMOTE, Random Oversampling, and Random Undersampling. Visualization of results was done with Matplotlib and Seaborn, distribution of the datasets, balancing of the results and performance measures of the classifiers.

These libraries were integrated to enable smooth interoperability among the pre-processing of data, feature engineering, training models and evaluation of model performance, and created a unified pipeline that could be used to process vast Arabic textual data.

4.2.3 Experimental Evaluation Strategy

Two complementary strategies were used to measure model performance to make sure it was robust and reduce overfitting. Initially, the cleaned labeled dataset was partitioned using an 80%–20% train–test split, where 80% of the data was allocated for model training and the remaining 20% was reserved as an unseen test set to evaluate generalization performance. Second, a 5-fold cross-validation procedure was conducted on the training set, whereby the data were partitioned into five equal subsets. In each iteration, the model was trained on four subsets and validated on the remaining one, ensuring that every sample was used for validation exactly once.

A 5-fold configuration was selected instead of 10-fold cross-validation to balance computational efficiency and reliable performance estimation, particularly given the size of the dataset and the training cost of deep learning and transformer-based models.

This is done in a cyclical way to minimize bias caused by a single train test split, and gives more accurate performance estimates by averaging between a series of validation runs. The train-test

split as well as cross-validation also make sure that the evaluation is statistically stable and efficient in terms of computing power.

4.2.4 Evaluation Metrics

Four traditional performance measures, including Accuracy, Precision, Recall and F1-Score, were used to measure effectiveness of the sentiment classification models. These measures are quantitative indicators of various facets of performance of classifier that are commonly used in supervised NLP evaluation. In addition, the Confusion Matrix and the Receiver Operating Characteristic (ROC) curve were used to provide a more detailed performance analysis.

- **Accuracy**

The measure of accuracy is the ratio of all the correctly classified samples to the overall sample. It is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.1)$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives.

- **Precision**

Precision measures how the model can accurately give positive sentiment without getting negative examples of sentiment correctly identified. It is formulated as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4.2)$$

Higher precision indicates lower false-positive rates.

- **Recall**

The recall (or sensitivity) is the ratio of the number of correctly determined positive samples by the model:

$$Recall = \frac{TP}{TP+FN} \quad (4.3)$$

Low recall indicates many missed positive samples.

- **F1-Score**

The F1-Score is a combination of the two metrics; precision and recall, which is measured using their harmonic mean:

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

- **Confusion Matrix**

The Confusion Matrix is a tabular representation of classification outcomes that summarizes the number of true positives, true negatives, false positives, and false negatives. It provides a detailed breakdown of prediction errors and allows for direct interpretation of how the model confuses one class with another.

- **ROC Curve**

The Receiver Operating Characteristic (ROC) curve evaluates the trade-off between True Positive Rate (Recall) and False Positive Rate across different classification thresholds. The Area Under the Curve (AUC) provides a threshold-independent measure of separability, indicating the model's ability to distinguish between positive and negative classes. The ROC analysis was originally introduced in signal detection theory and later popularized in machine learning evaluation by Fawcett (2006), who formalized its application for classifier performance assessment.

In addition to the above metrics, statistical hypothesis testing is employed to assess whether observed differences in model performance are statistically significant. Specifically, McNemar's test is applied to compare the predictions of paired classifiers on the same test set. McNemar's test evaluates whether two models make errors on the same instances or on different ones, using a chi-squared statistic computed from the discordant prediction pairs. A p-value below 0.05 indicates

that the performance difference between two models is statistically significant and not attributable to chance. This test is particularly suitable for binary classification tasks and provides a rigorous basis for concluding that one model genuinely outperforms another, beyond what raw accuracy or F1-score comparisons alone can establish.

4.3 Evaluation of Annotation Consistency Across Labeling Techniques

Once the process of data cleaning and normalization was done, the completed dataset was subjected to a multi-source labeling to test the consistency and reliability of the sentiment annotation. Three methods of labeling were considered: manual labeling done by a native Arabic speaker who has domain knowledge, the labeling by the GPT-4o model and rating-based labeling based on the Google Play review scores. The aim of the comparison of these methods was to establish the degree to which the human interpretation, algorithmic inference and rating heuristics are in agreement in assigning sentiment categories to reviews of Arabic banks. Consistency analysis, which was summarized in Table 4.1, proved a slight difference between the three methods. Manual and rating-based labeling showed the middle point of not less than 70.7 percent of the agreement that numerical ratings do not necessarily reflect the tone of the material of the review text, especially when users write short, neutral descriptions accompanied by a strongly positive or negative rating. GPT-4o and rating-based labeling agreement was slightly lower, and this was 66.5, which serves as evidence that automated sentiment interpretation varies more often than score-based labeling which supports the constraints of assuming a direct relationship between star ratings and textual sentiment. The best accuracy of 88.1 was achieved between manual and the GPT-4o and this indicates that the model can capture sentiment cues that are much closer to human linguistic intuition. However, an unresolved discrepancy of 11.9 percent shows that there are still ongoing issues and of particular concern is the variability in the presence of subtle sentiment expression, sarcasm, cultural idioms, dialectal variety, and review mixes of positive and negative statements. The results highlight the fact that despite GPT-4o being able to effectively simulate human-level annotation performance on the sentiment tasks in Arabic, expert human judgment is still obligatory whenever operating on a more sophisticated linguistic phenomenon. Therefore, hybrid labeling approaches that combine the benefits of expert labeling and automated

help can offer the most credible basis of sentiment classification in sophisticated, dialect-laden areas of the Arabic banking app.

Table 4.1: Agreement and Difference Rates Between Sentiment Labeling Techniques

Comparison Method	Agreement Rate	Difference Rate
Manual vs. Rating	70.7%	29.3%
GPT-4o vs. Rating	66.5%	33.5%
Manual vs. GPT-4o	88.1%	11.9%

4.4 Evaluation Model

To identify the effectiveness of the proposed SA architecture, a series of experiments were implemented in terms of conventional ML classifiers, DL models, and the large language models (LLM). This multi-level test enables drawing parallels between the multimodal analytical strategy, the application of neural networks, and the generation models of the most recent generation when it comes to processing reviews of Arabic banks. The further parts, introduce and discuss performances of these models basing on standardized measures, their advantages, weaknesses, and suitability to real-life SA of tasks within the banking industry.

4.4.1 Evaluation of Machine Learning Models

The comparison of the ML models between the training and testing sets gives the explicit idea about the response of various classifiers to textual sentiment features and balancing strategies when used in Arabic banking reviews. Table 4.2 summarizes the results of all the models which were tested on the remaining 20 percent of the data after being trained on the initial 80% of the data using a binary sentiment scheme so as to give a consistent and fair comparison of the performance of all models. In all the experiments, the most stable and precise classifiers were the Logistic Regression and Support Vector Machine (SVM) as they had the greatest accuracy and F1-scores across all balancing conditions. Logistic Regression scored an accuracy of 0.9260 and F1-score of 0.9111 when applying class weights to correct imbalance on the training, just behind SVM which scored 0.9239 accuracy and 0.9092 F1-score. Both Logistic Regression and SVM produced

high recall values of 0.9250 and 0.9302, respectively, indicating their strong ability to correctly identify positive sentiment instances in the unseen test data. High recall in this context suggests that the models successfully capture discriminative sentiment patterns learned during training and are able to generalize them to new data without missing a substantial number of true positive cases. This reflects effective decision boundary formation and reduced underfitting, particularly important in banking sentiment analysis where failing to detect positive or negative user signals may impact interpretation accuracy.

Both models were consistent when tested on SMOTE augmentation with Logistic Regression returning 0.9258 accuracy and 0.9109 F1-score and SVM returning 0.9241 accuracy and 0.9092 F1-score. The almost similar outcomes of class weighting and SMOTE imply that such linear models are not as much concerned with methods of data balancing because they are able to build clear separation boundaries in high-dimensional TF-IDF spaces. Conversely, it can be argued that Random Forest always performed the worst in both the training and test stage, particularly with the most accurate class weighting of 0.8469, F1-score of 0.8273 and SMOTE accuracy of 0.8516, F1-score of 0.8316, probably because it is more prone to overfitting and struggles with the sparse feature of the text. Naive Bayes had moderate scores on all balancing techniques, and was found to perform better with oversampling since it relies on word-frequency independence, which has the advantage of more and diversified minor class samples.

Interestingly, the precision, recall, and F1-scores of Logistic Regression and SVM remained highly consistent under both oversampling and undersampling strategies, demonstrating their robustness when the training distribution is artificially modified.

Under the oversampling approach, slightly improved performance was observed. SVM achieved an accuracy of 0.9258, while Logistic Regression reached 0.9239, indicating a marginal reduction in false positive predictions in positive sentiment classification.

Similarly, the undersampling strategy yielded comparable results. Logistic Regression attained an accuracy of 0.9256 with an F1-score of 0.9258, whereas SVM achieved an accuracy of 0.9224 and an F1-score of 0.9226. These results suggest that both models maintain strong discriminative capability even when trained on reduced datasets, confirming their ability to learn stable and meaningful sentiment representations despite variations in training sample distribution.

These results of the training and testing show that Logistic Regressions and SVM are most appropriate in the Arabic sentiment classification because they always perform better than other classifiers despite the balancing method used. The fact that they generalize well in both training and testing sets proves their applicability in large-scale, high-dimensional textual data with semantic nuances and dialectal variations demanding models with the sensitivity to subtle sentiment boundaries. These results put in high emphasis the consideration of the selection of correct classifiers to Arabic banking data and show that linear models are one of the most promising options of SA when used in practice.

Table 4.2: Machine Learning Performance Under Different Balancing Techniques

Model	Balancing Method	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	Class Weights	0.8914	0.9048	0.8215	0.8612
SVM	Class Weights	0.9239	0.8892	0.9302	0.9092
Logistic Regression	Class Weights	0.9260	0.8976	0.9250	0.9111
Random Forest	Class Weights	0.8469	0.7695	0.8945	0.8273
Naïve Bayes	SMOTE	0.8883	0.8716	0.8531	0.8622
SVM	SMOTE	0.9241	0.8916	0.9276	0.9092
Logistic Regression	SMOTE	0.9258	0.8967	0.9255	0.9109
Random Forest	SMOTE	0.8516	0.7773	0.8939	0.8316
Naïve Bayes	Oversampling	0.8874	0.8873	0.8874	0.8873
SVM	Oversampling	0.9245	0.9258	0.9245	0.9248
Logistic Regression	Oversampling	0.9239	0.9239	0.9239	0.9239
Random Forest	Oversampling	0.8465	0.8556	0.8465	0.8476

Naïve Bayes	Undersampling	0.8849	0.8848	0.8849	0.8848
SVM	Undersampling	0.9224	0.9236	0.9224	0.9226
Logistic Regression	Undersampling	0.9256	0.9264	0.9256	0.9258
Random Forest	Undersampling	0.8471	0.8574	0.8471	0.8483

The implementation of the ML models was further justified and generalized as a 5-fold cross-validation process was done on the binary sentiment data. The approach offers a more confident approximation of the predictive capability in the real world since it trains and tests each model five times on five data partitions, which reduces the possibility of bias due to one train / test split. The results of the cross-validation as it is demonstrated in Table 4.3 support the tendencies of the first analysis: Logistic Regression and Support Vector Machine (SVM) have the best results in all balancing strategies. Logistic Regression performed better than SVM with 0.9256 average accuracy and 0.9099 F1-score when SMOTE was used, which is consistent with the stable behavior of these algorithms with the presence of synthetic minority samples. Naïve Bayes, which was performing fairly, exhibited smaller F1-scores in the range of 0.868 to 0.894, which suggests that it is less able to capture more complex sentiment patterns, especially when the data contains linguistic ambiguity and dialectal variability as is the case in Arabic user reviews.

Equally, as noted by the results of the oversampling experiments, the resilience of Logistic Regression and SVM is repeated, with the pair of algorithms scoring almost the same values in both accuracy between 0.924-0.925 and F1-score between 0.923-0.925, demonstrating that the two algorithms still outperform their counterparts despite being fed with more records of the minority class by duplication, rather than synthesis. Naive Bayes was stable with oversampling with a lower level of performance of 0.894 and F1, whereas Random Forest still lagged behind all the other models, as its top accuracies were not over 0.850, no matter the balancing method. The trend continues in the context of undersampling whereby Logistic Regression attained 0.9243 accuracy and 0.9243 F1, and SVM attained 0.9214 accuracy and 0.9216 F1. The small values of standard deviations across the folds are generally between 0.0019 and 0.0044, which ensure the good consistency and stability of the linear models. Conversely, the higher variability of the Random Forest with maximum standard deviation values of 0.0068 indicates its sensitivity to smaller data

and the inability to model sparse and high-dimensional TF-IDF vectors. The 5-fold cross-validation results indicate that Logistic Regression and SVM can be used to provide a strong and reliable performance on Arabic sentiment classification, and Naïve Bayes can be used as a useful, but inaccurate baseline, and Random Forest cannot be used as an efficient model, unless additional feature engineering efforts and model adaptation are made.

Table 4.3: Machine Learning Cross-Validation Results Under Different Balancing Techniques

Model	Balancing Method	Accuracy (Mean)	Accuracy (Std)	F1-Score (Mean)	F1-Score (Std)
Logistic Regression	SMOTE	0.925609	0.001998	0.909954	0.002837
SVM	SMOTE	0.922385	0.002814	0.906354	0.003838
Naïve Bayes	SMOTE	0.894266	0.003332	0.868878	0.004486
Random Forest	SMOTE	0.849860	0.005518	0.829294	0.006800
Logistic Regression	Oversampling	0.924930	0.002626	0.924954	0.002678
SVM	Oversampling	0.923827	0.001941	0.923945	0.001949
Naïve Bayes	Oversampling	0.894181	0.003220	0.894003	0.003261
Random Forest	Oversampling	0.849012	0.004741	0.850132	0.004721
Logistic Regression	Undersampling	0.924251	0.002109	0.924283	0.002216

SVM	Undersampling	0.921410	0.002231	0.921555	0.002244
Naïve Bayes	Undersampling	0.894096	0.003206	0.893968	0.003256
Random Forest	Undersampling	0.844983	0.004441	0.846148	0.004397

4.4.2 Evaluation of Deep Learning Models

In order to explore the effects of the DL on the Arabic sentiment classification, an LSTM model was trained and tested on the same binary dataset with the use of SMOTE and random oversampling to manage the imbalance between classes during the training process. In contrast to the conventional ML based classifiers, LSTM is explicitly suited to capture sequential dependencies and contextual links in the text hence is perfect when it comes to addressing the morphological richness and dialectal diversity inherent in Arabic user reviews.

4.4.2.1 Evaluation of LSTM Model Using SMOTE

In order to assess the performance of DL on Arabic sentiment classification, an LSTM model was trained with the help of SMOTE to balance the binary training data and early stopping was used to prevent overfitting in the process of training. The early stopping mechanism monitored the validation loss and halted training when performance no longer improved, thereby preventing the model from overfitting to oversampled synthetic data. Figure 4.1 shows the development of training and validation accuracy with the epochs, which reveals that the training accuracy has improved significantly, as the training accuracy of 0.74 in the first epoch has increased to a high of 0.98 in the fifth epoch, but the validation accuracy has remained at 0.94 in the initial epochs and then decreased slightly after the second epoch. On the same note, the validation loss curve started reducing but started to increase after the third epoch, which means that the overfitting has started. This ratio of training to validation performance supports the significance of early stopping as a regularization technique, which guarantees that the model learns meaningful sentiment features and not noise produced by artificial samples created using SMOTE.

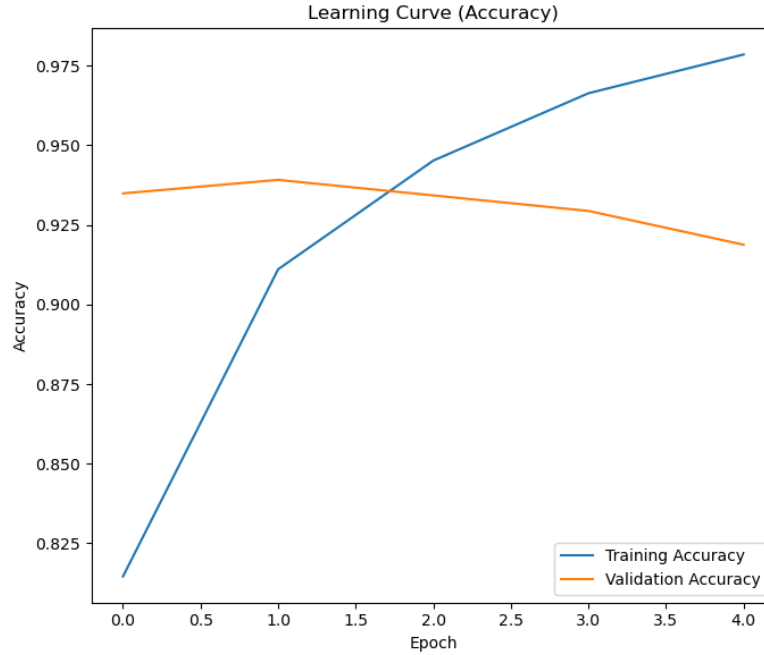


Figure 4.1: Training and Validation Accuracy of LSTM

The classification results presented in Table 4.4 demonstrate that the LSTM model achieved strong performance in binary Arabic sentiment classification, with an overall accuracy of 0.94 and balanced macro- and weighted-average F1-scores of 0.94.

For the negative class, the model obtained a precision of 0.98 and a recall of 0.92, resulting in an F1-score of 0.95. This indicates that when the model predicts a review as negative, it is highly reliable (very low false-positive rate), although a small proportion of actual negative reviews are not captured (moderate false negatives). For the positive class, the model achieved a precision of 0.89 and a recall of 0.97, with an F1-score of 0.93. This suggests that the model is highly effective at identifying positive reviews (very low false-negative rate), though it occasionally misclassifies some negative samples as positive.

The slight difference in precision between the two classes indicates that the model is more conservative when predicting the negative class and slightly more permissive when predicting the positive class. From an application perspective, a stronger precision in the negative class is preferable in banking sentiment analysis, as it ensures that reviews classified as dissatisfaction are indeed genuine complaints, reducing the risk of falsely flagging satisfied customers as dissatisfied.

The balanced macro and weighted F1-scores indicate stable and fair performance across both classes. These results show that the LSTM effectively captures contextual and sequential patterns in Arabic text, while SMOTE enhances minority-class learning without harming generalization.

Table 4.4: Classification Report of the LSTM Model Trained with SMOTE for Binary Arabic Sentiment Classification

Class	Precision	Recall	F1-Score	Support
Negative	0.98	0.92	0.95	2783
Positive	0.89	0.97	0.93	1933
Accuracy	0.94			4716
Macro Avg	0.93	0.94	0.94	4716
Weighted Avg	0.94	0.94	0.94	4716

In order to investigate predictive behavior further, Figure 4.2 shows the confusion matrix of the LSTM model. The model correctly classified 2,553 out of 2,783 negative reviews (91.7%) and 1,876 out of 1,933 positive reviews (97.1%). Misclassifications were limited to 230 negative reviews predicted as positive and 57 positive reviews predicted as negative.

This distribution confirms the strong recall observed for the positive class and the high precision for the negative class reported in the classification report. The matrix indicates balanced sensitivity and specificity, demonstrating that the model effectively detects positive sentiment while remaining reliable in identifying dissatisfied users—an important consideration in banking applications where negative feedback carries strategic significance.

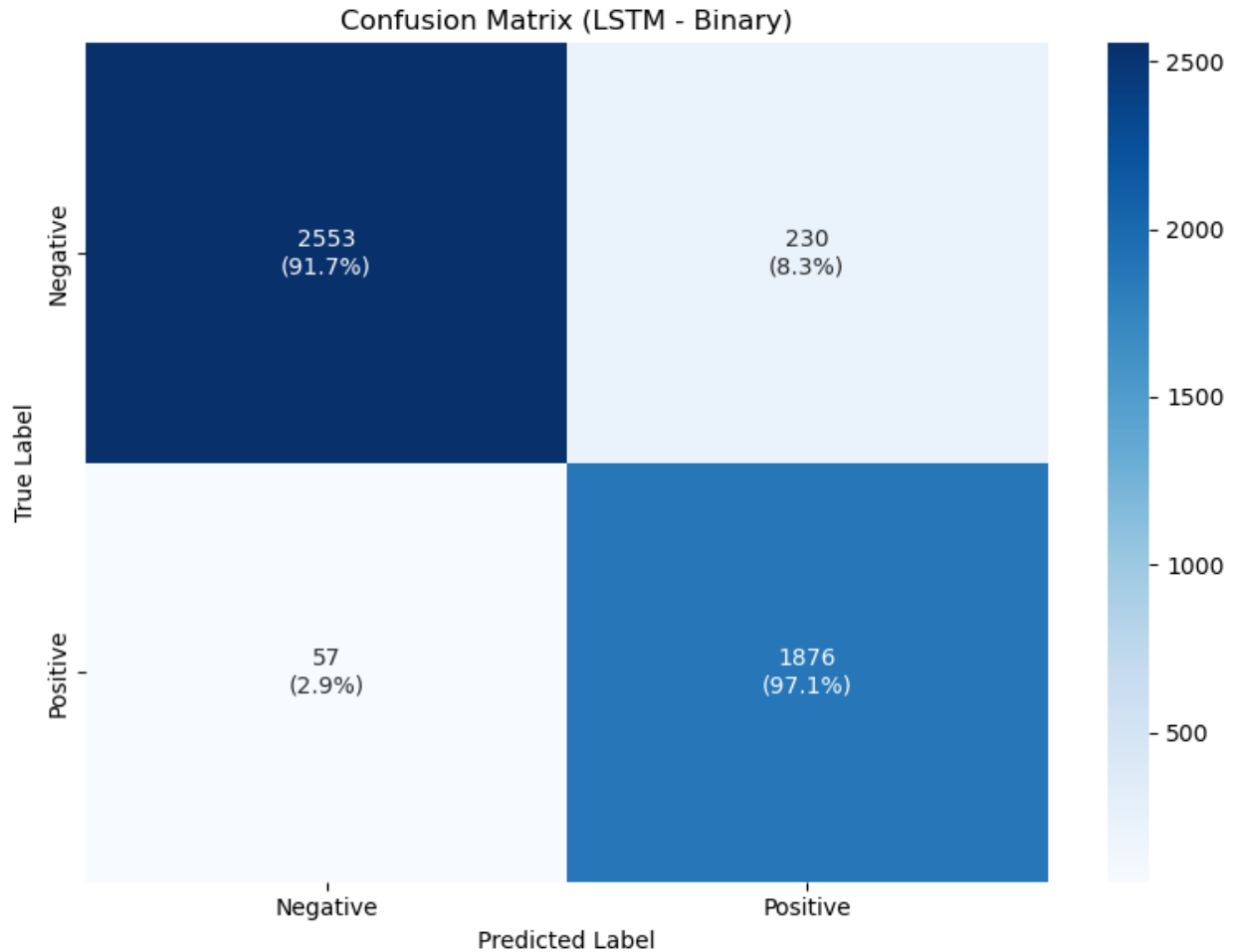


Figure 4.2: Confusion Matrix of the LSTM Model with SMOTE

Lastly, the LSTM high discriminative ability is also justified by the ROC curve in Figure 4.3 with AUC of 0.99. The sharp rise of the curve towards the upper-left suggests a very high true positive rate at very low false positive rates. This finding supports the fact that this model works effectively with different set of classification thresholds, implying great robustness to real-time or automated deployment use.

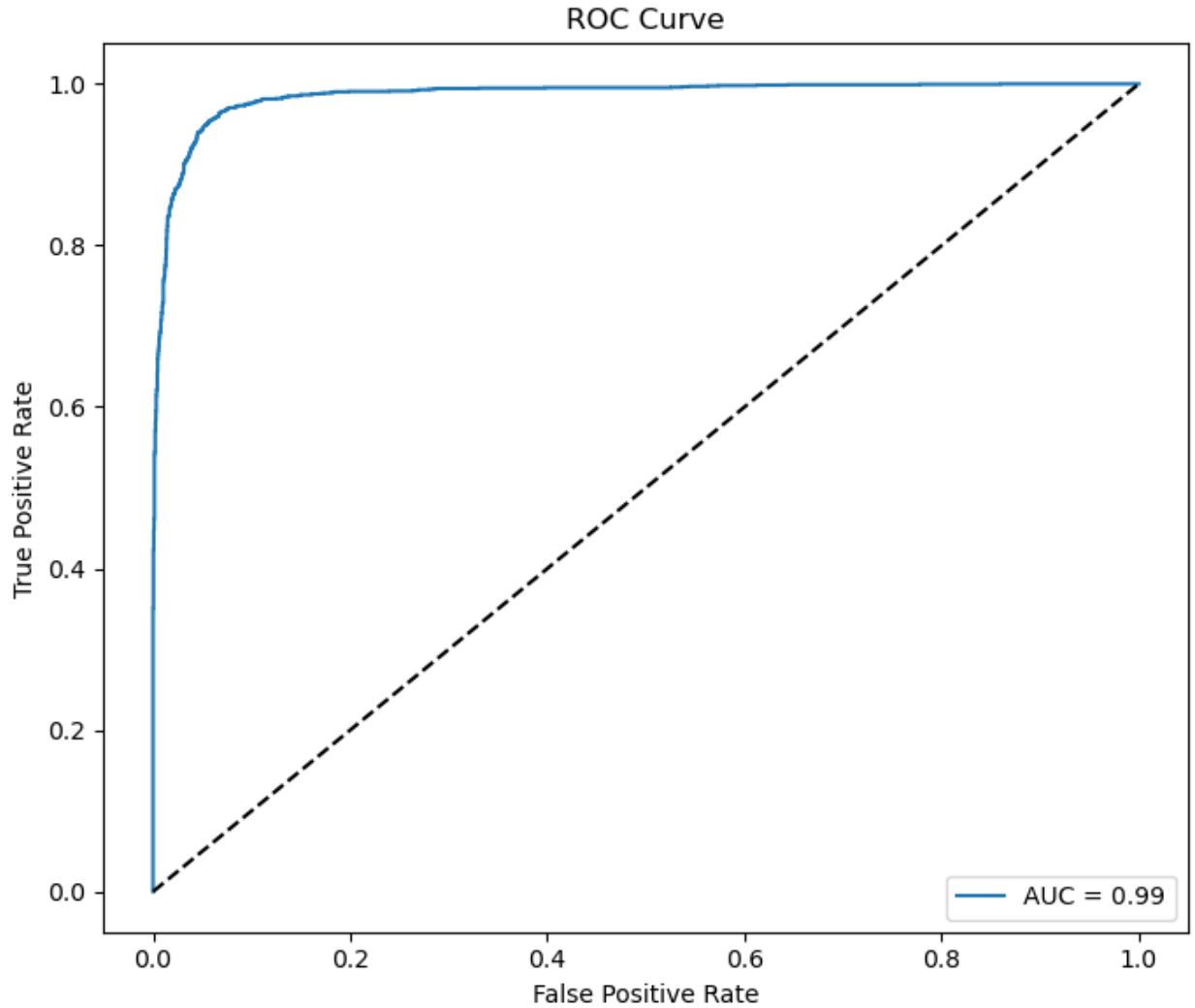


Figure 4.3: ROC Curve of the LSTM Model with SMOTE

The accuracy of the model shown through evaluation metrics and graphical diagnostics prove that the LSTM model is one of the viable DL methods of SA in Arabic banking systems and can address the subtleties of emotion expression and dialect patterns in a more effective way than classical models.

4.4.2.2 Evaluation of LSTM Model Using Oversampling

To determine the effectiveness of DL based on various balancing strategies, a second LSTM model was trained with random oversampling to balance the proportion between positive and negative sentiment classes. Like in the above experiment, an early stopping criterion was used to evaluate the loss on validation loss in training to ensure that the model is not overfitting on duplicated and

oversampled minority samples. Figure 4.4 shows the correlation between training and validation accuracy and loss at seven epochs. The training accuracy also improved quickly with the initial epoch of 0.77 and after 5 epochs to above 0.99 with a significant reduction in the loss of 0.46 to 0.02. In the meantime, the maximum of the validation accuracy was achieved at 0.98 between the third and fifth epochs, whereas the validation loss peaked at the middle of the training process and then increased more slowly after the sixth epoch. This deviation of training and validation performance is an indication of the onset of overfitting thus showing the need to use early stopping in order to ensure the best generalization in the learning of duplicated samples.

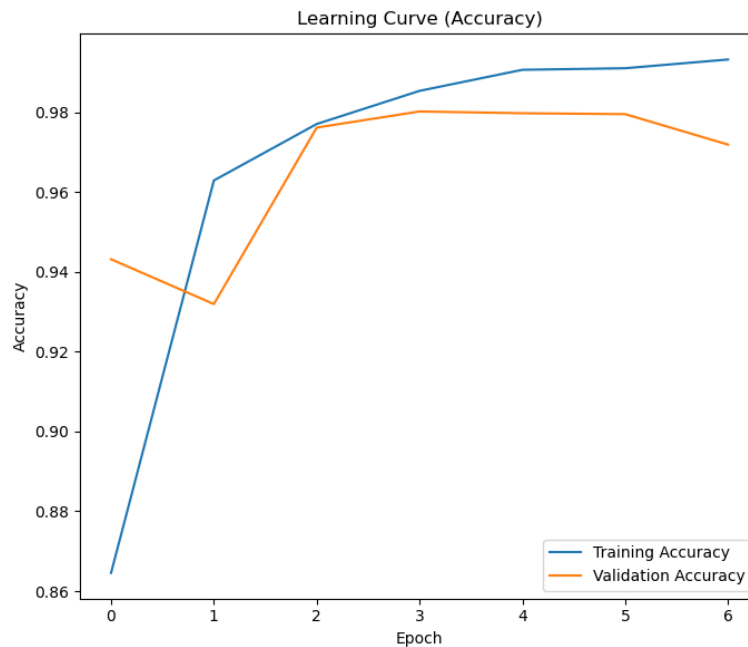


Figure 4.4: Training and Validation Accuracy Using Random Oversampling

The performance metrics as they are summarized in Table 4.5 also help to prove that the oversampled LSTM model is effective in classifying Arabic sentiment. The model had the overall accuracy of 0.95, and F1-scores of 0.95 and 0.94 in negative and positive reviews, respectively, which suggests that the quality of the predictions is similar in both categories. The accuracy and recall rates were also significant and were 0.97 precision and 0.94 recall in the negative and 0.92 precision and 0.96 recall in the positive class. These measures indicate sampling succeeds in improving coverage of the minority group without negatively affecting accuracy, which allows the LSTM to detect subtle features of sentiment in both categories.

Random oversampling yielded slightly better results than SMOTE because it preserves the original linguistic structure of minority-class samples without introducing synthetic variations. In short-text Arabic reviews, where expressions are highly contextual and dialect-sensitive, artificially generated samples (as in SMOTE) may distort subtle semantic patterns. In contrast, simple duplication of real minority examples reinforces authentic sentiment signals, enabling the LSTM to better learn genuine contextual and sequential patterns, which results in improved stability and overall classification performance.

Table 4.5: Classification Report of the LSTM Model Trained with Random Oversampling for Arabic Sentiment Classification

Class	Precision	Recall	F1-Score	Support
Negative	0.97	0.94	0.95	2783
Positive	0.92	0.96	0.94	1933
Accuracy	0.95			4716
Macro Avg	0.94	0.95	0.94	4716
Weighted Avg	0.95	0.95	0.95	4716

The confusion matrix in Figure 4.5 given further insight into model behavior and shows that 93.9% of negative reviews and 95.7% of positive reviews have been correctly identified. The percentage of misclassification was also very low, and only 171 false negative and 84 false positive reviews were predicted. The lower number of false negatives (171) compared to SMOTE indicates improved detection of actual positive reviews, meaning fewer truly positive cases were incorrectly classified as negative. This reduction reflects stronger minority-class sensitivity and a more reliable decision boundary when using random oversampling. This balance is an indicator of a strong decision boundary that has been learnt by the model and it confirms the success of oversampling in increasing sensitivity to minority-classes.

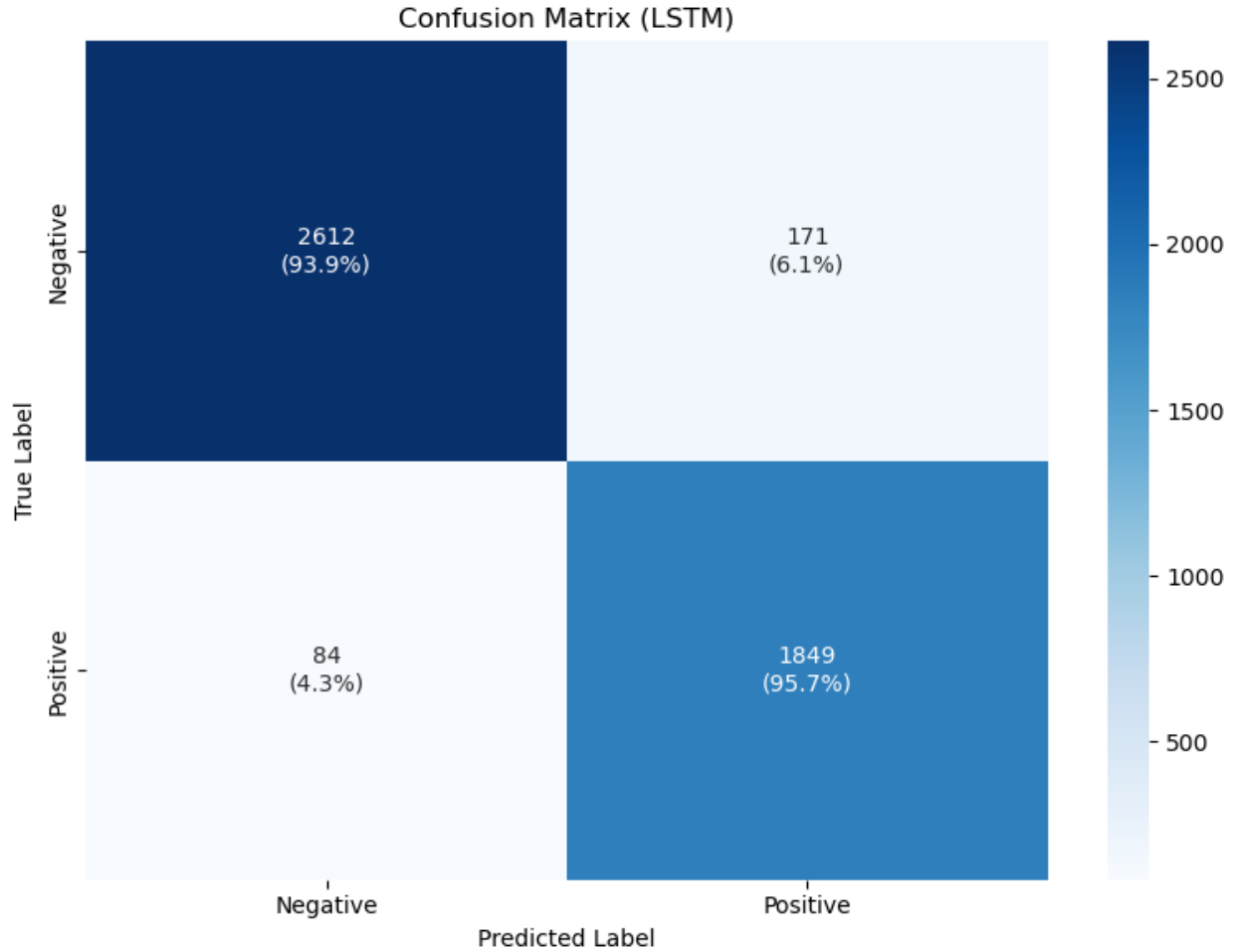


Figure 4.5: Confusion Matrix of the LSTM Model for Binary Sentiment Classification using Random Oversampling

Lastly, Figure 4.6 has shown that the ROC curve has an Area Under the Curve (AUC) score of 0.98 which reflects outstanding discriminative capability at different classification thresholds. The sharp curve towards the upper-left corner validates the ability of the model to reduce false positives at the expense of consistent recall which supports the usefulness of the model in large scale SA in banking feedback systems. Combined, these results of the evaluation indicate that LSTM with oversampling not only preserves high accuracy and consistency but also enhances the balance of classes making it a highly competitive DL model to Arabic sentiment classification.

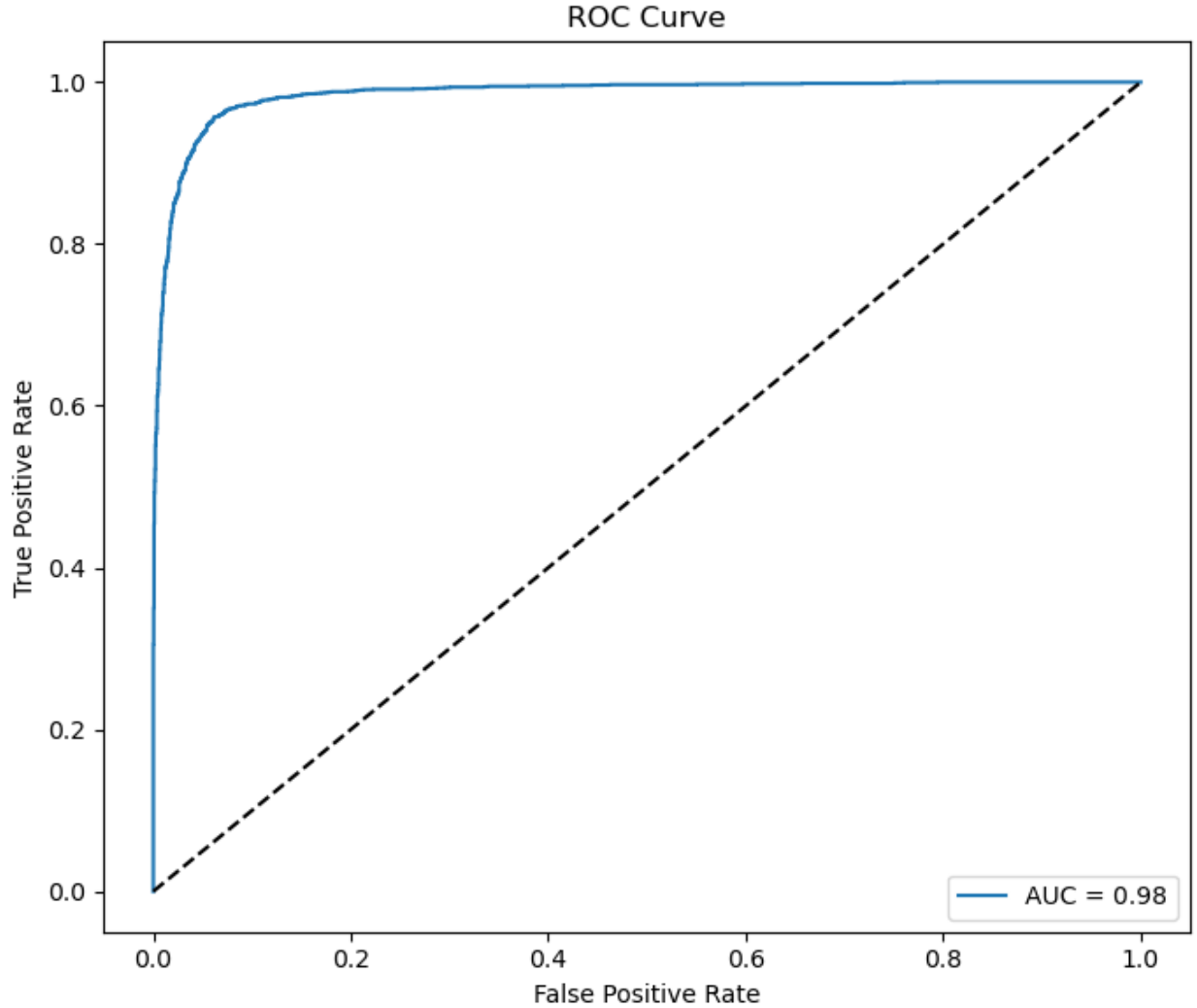


Figure 4.6: ROC Curve of the LSTM Model for Binary Sentiment Classification using Random Oversampling

An evaluation of the two LSTM balancing strategies comparatively shows that the two methods, SMOTE and random oversampling, are both effective in improving sentiment classification but have their minor differences in the model behavior. The LSTM with SMOTE performed well in generalization, where its accuracy equaled to 0.94 with less cases of overfitting because it has created artificial minority samples that enhance diversity of features space. Conversely, the oversampling-based LSTM was slightly more accurate of 0.95 and minority-class recall, since more positive samples with duplications were exposed. However, random oversampling exhibited earlier onset of overfitting in the form of steep training gains and validation stagnation. These findings suggest that both methods are efficient, but SMOTE is more regularised and has more

feature diversity as opposed to oversampling which has slightly higher accuracy but has less overfitting control.

4.4.3 Evaluation of LLM Models

In a further attempt to support the usefulness of transformer-based architectures to Arabic sentiment classification, two dedicated large language models, AraBERT and MarBERT, were fine-tuned on the annotated dataset of Arabic reviews in mobile banking. The two models are trained on large Arabic corpora and are produced to be better at capturing morphological richness, dialects and contextual nuances than the traditional ML and recurrent neural models. Their comparison gives a better insight into how domain-independent pretraining can be transferred to domain-specific sentiment tasks to enable a solid comparison to classical and DL strategies.

4.4.3.1 Evaluation of AraBERT Model

To investigate the contribution of transformer-based architectures to Arabic SA, the AraBERT model was fine-tuned on the balanced dataset using class weights to mitigate bias toward the majority class. By integrating class weights into the loss function, the training process penalized misclassification of underrepresented positive samples more heavily, thereby enforcing equitable learning across sentiment categories.

Figure 4.7 illustrates the training and validation loss curves of the AraBERT model across epochs. The training loss decreases consistently from epoch 1 to epoch 4, indicating effective learning and progressive optimization of model parameters. In contrast, the validation loss decreases initially and reaches its minimum around epoch 2, after which it begins to increase. This divergence between training and validation loss suggests the onset of overfitting beyond the second epoch, where the model continues to fit the training data more closely while its generalization performance on unseen data slightly deteriorates. This behavior highlights the importance of early stopping to retain the model at its optimal generalization point.

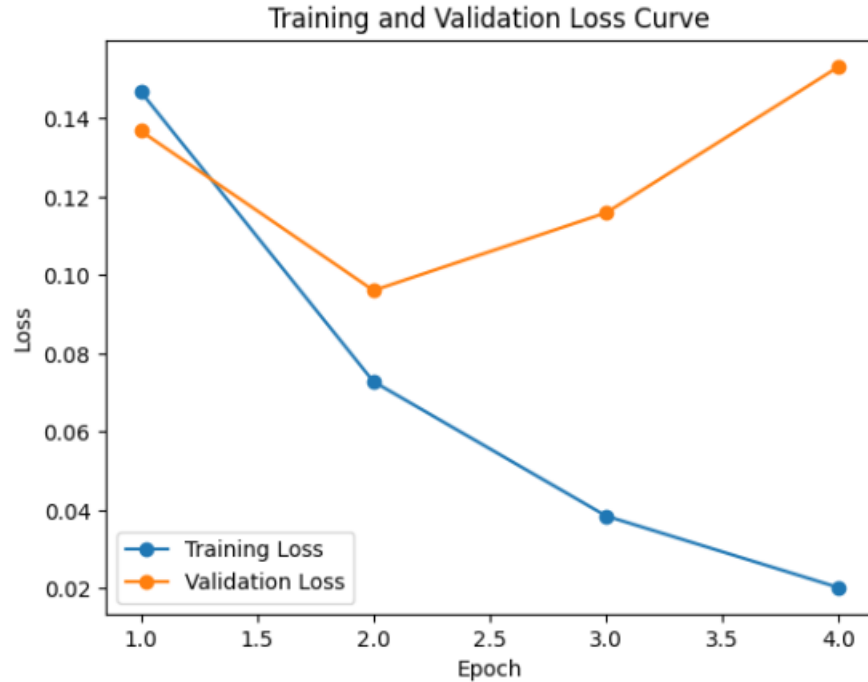


Figure 4.7: Training and Validation Loss of AraBERT

Table 4.6 shows training and validation performance change throughout four epochs. The model was convergence speedy and after a few training losses in epoch 1, loss steadily dropped to 0.0202 in the fourth epoch and the other way round, which signified an optimal learning plateau. Validation accuracy was stable, with a score of above 0.97 upon the second epoch, and precision-recall results indicated balanced learning, with F1-scores in the range of 0.96 and 0.97, indicating that the model was good at both learning the salient linguistic cues and not overfitting.

Table 4.6: Training and Validation Performance of AraBERT

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.146800	0.136782	0.967557	0.964995	0.955510	0.960229
2	0.072800	0.096076	0.973707	0.955209	0.981893	0.968367
3	0.038500	0.115919	0.974979	0.956259	0.983963	0.969913
4	0.020200	0.153241	0.973919	0.960794	0.976203	0.968437

The last performance indicated in Table 4.7 shows the high generalization capacity of AraBERT in the unseen samples. The general accuracy of the model was 0.97, the negative sentiment classification had F1-score of 0.98, and the positive one was 0.97. The bad class showed very high accuracy of 0.99 which is the ability of the model to identify the bad users with the least false alarms, whereas the good class showed a better recall of 0.98 which means that the model identified favorable customer feel strongly. These findings highlight the efficiency of AraBERT in dealing with complicated morphology and contextual variations of the Modern Standard Arabic language that are found in the context of mobile banking reviews.

Table 4.7: Classification Report of AraBERT Model

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.99	0.97	0.98	2783
Positive (1)	0.96	0.98	0.97	1933
Accuracy	0.97			4716
Macro Avg	0.97	0.97	0.97	4716
Weighted Avg	0.97	0.97	0.97	4716

The observations are further supported by Figure 4.8, which gives the confusion matrix, which shows equal distribution of errors between the two sentiments. The model was able to recognize 2694 negative reviews and 1898 positives review correctly; 89 negative samples were misclassified as positives and 35 positive samples were misclassified as negative. The presence of very low misclassification rates in this distribution proves that AraBERT is better suited to discriminate and offer greater reliability than the more traditional ML and recurrent neural models. Altogether the findings confirm that class weighting, used together with a strong Arabic language model like AraBERT, provides a strong sentiment classifier that can help distinguish subtle semantic differences that are common in user-generated financial reviews.

The relatively higher misclassification rate for the positive class can be attributed to the linguistic complexity of Arabic banking reviews. Users frequently embed negative sentiment within polite

or formal expressions, and dialectal variations in expressing satisfaction or dissatisfaction create ambiguity that the model resolves in favor of the majority training distribution. Additionally, short positive reviews with minimal contextual cues provide less discriminative information for the model compared to longer, more explicit negative reviews.

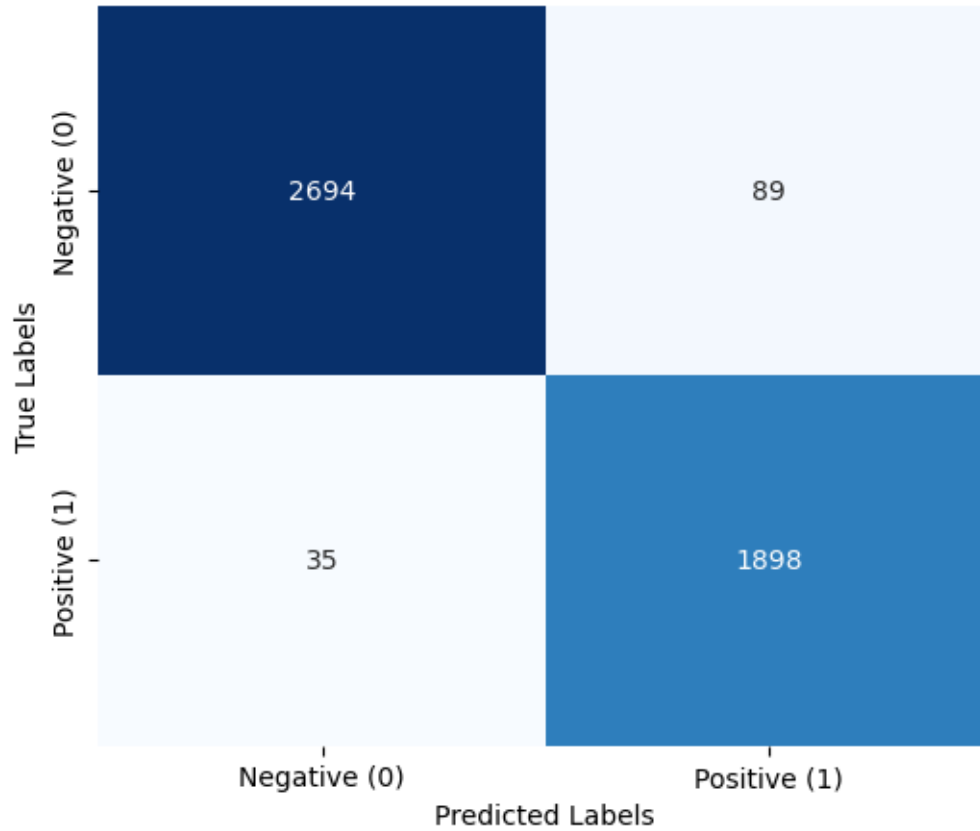


Figure 4.8: Confusion Matrix of AraBERT Model for Binary Sentiment Classification

4.4.3.2 Evaluation of MarBERT Model

The MarBERT model was fine-tuned on the sentiment-labeled mobile banking datasets so as to explore the effectiveness of advanced transformer architectures on dialectal Arabic using the weighted class with class imbalance. In contrast to AraBERT, which is initially pretrained mostly on Modern Standard Arabic, MarBERT has the advantage of large exposure to regional dialects as often observed in informal user reviews in online environment.

Figure 4.9 presents the training and validation loss curves of the MARBERT model across four epochs. The training loss decreases consistently, indicating effective parameter optimization and

stable learning. The validation loss shows a sharp decline from epoch 1 to epoch 2 and stabilizes at epoch 3, before slightly increasing at epoch 4. This pattern suggests that the model achieves optimal generalization around epochs 2–3, after which mild overfitting begins to appear. The behavior confirms the importance of early stopping to preserve the best-performing model configuration.

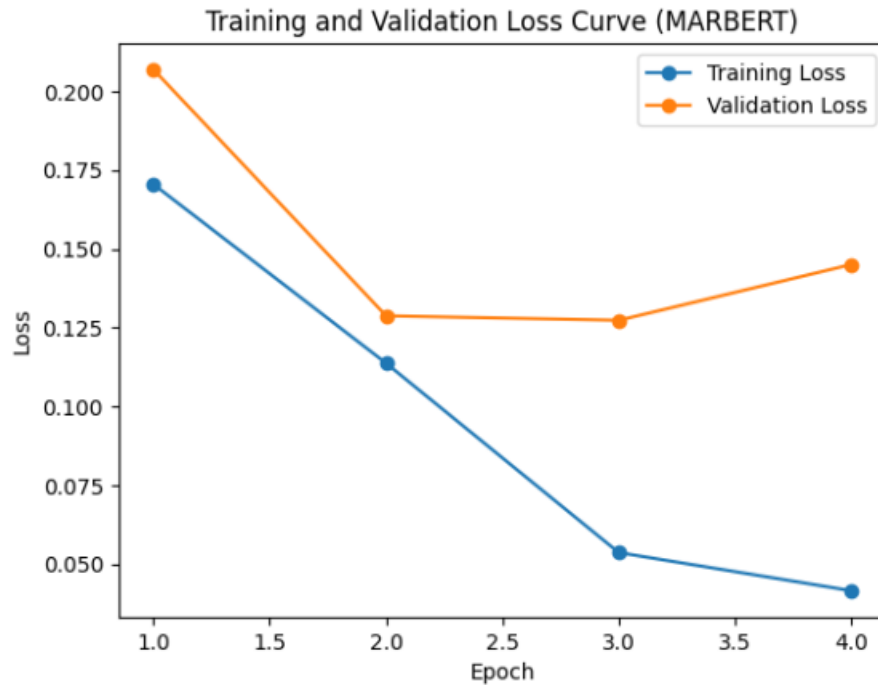


Figure 4.9: Training and Validation Loss of MarBERT

The results of the training and validation performance are summarized in table 4.8 after four epochs. The model was shown to reduce training loss, which started with a loss of 0.1705 at the first epoch and decreased down to 0.0415 at the 4th epoch, and the validation accuracy steadily rose at the end of the model to 0.9697. Validation loss experienced some small variations, but the precision, recall, and F1-scores were always sufficiently high, which means that training was effective and there was no drastic overfitting. Interestingly, the recall in the early epochs was higher than accuracy, indicating that MarBERT has a high capability to detect positive sentiment cues during an early phase, which could be because of its dialect-enhanced pretraining.

Table 4.8: Training and Validation Performance of MarBERT

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.170500	0.206914	0.958227	0.975877	0.920848	0.947565
2	0.113700	0.128733	0.963316	0.932220	0.981893	0.956412
3	0.053600	0.127314	0.969678	0.947053	0.980859	0.963659
4	0.041500	0.145026	0.969678	0.951109	0.976203	0.963492

Table 4.9 also demonstrates the performance of the final MarBERT model in terms of classification. The model predicted with an overall accuracy of 0.97, which is the same as AraBERT, but with slightly lower accuracy on the positive classification of 0.95, and a high recall of 0.98. This means that MarBERT is more likely to call some neutral or ambiguous reviews as positive because it is more interested in sentiment inclusiveness rather than boundary discrimination. The negative class demonstrated excellent accuracy of 0.99 and an F1-score of 0.97, which indicates the ability of the model to identify dissatisfaction with the least amount of false alarm. The weighted-average and macro-metrics meet at the same value of 0.97 indicating consistent and equal performance in the two sentiment categories.

From a banking operations perspective, the False Accept Rate (FAR) and False Reject Rate (FRR) provide operationally meaningful measures of classification errors. The FAR, defined as the proportion of negative reviews incorrectly classified as positive ($\text{FAR} = 1 - \text{Recall_Negative}$), equals $1 - 0.96 = 0.04$, or approximately 4.0% for MARBERT. The FRR, defined as the proportion of positive reviews incorrectly classified as negative ($\text{FRR} = 1 - \text{Recall_Positive}$), equals $1 - 0.98 = 0.02$, or approximately 2.0%.

In a banking context, minimizing FAR is of higher strategic priority, as undetected customer complaints (false acceptances) can lead to unresolved service failures, customer churn, and reputational damage. The FAR of 4.0% indicates that MARBERT correctly detects 96 out of every 100 genuine complaints, providing a high level of assurance for banking sentiment monitoring. Meanwhile, the low FRR of 2.0% ensures minimal false escalation of satisfied customer feedback,

preventing unnecessary resource allocation. Compared to traditional ML models, where SVM achieves a FAR of approximately 10% and an FRR of 11%, the transformer-based models demonstrate substantially lower error rates, further validating their suitability for deployment in financial service applications.

Table 4.9: Classification Report of MarBERT Model

Class	Precision	Recall	F1-Score	Support
Negative (0)	0.99	0.96	0.97	2783
Positive (1)	0.95	0.98	0.96	1933
Accuracy	0.97			4716
Macro Avg	0.97	0.97	0.97	4716
Weighted Avg	0.97	0.97	0.97	4716

The balanced classification ability of the model is confirmed by the confusion matrix of MarBERT in figure 4.10. The model rightly distinguished 2677 negative reviews and 1896 positive reviews with only 106 negative samples being classified as positive and 37 positive reviews being negative. These low misclassification rates support the idea that MarBERT is highly sensitive to the sentiment polarity even in the face of dialect-specific expressions, colloquialisms and informal writing structures common in mobile app review data. These findings testify to the fact that the combination of class weighting and dialect-aware language modeling provided by MarBERT produces a competitive and stable sentiment classifier, which can be used to assess the customer perceptions of Arabic digital banking systems on a large scale.

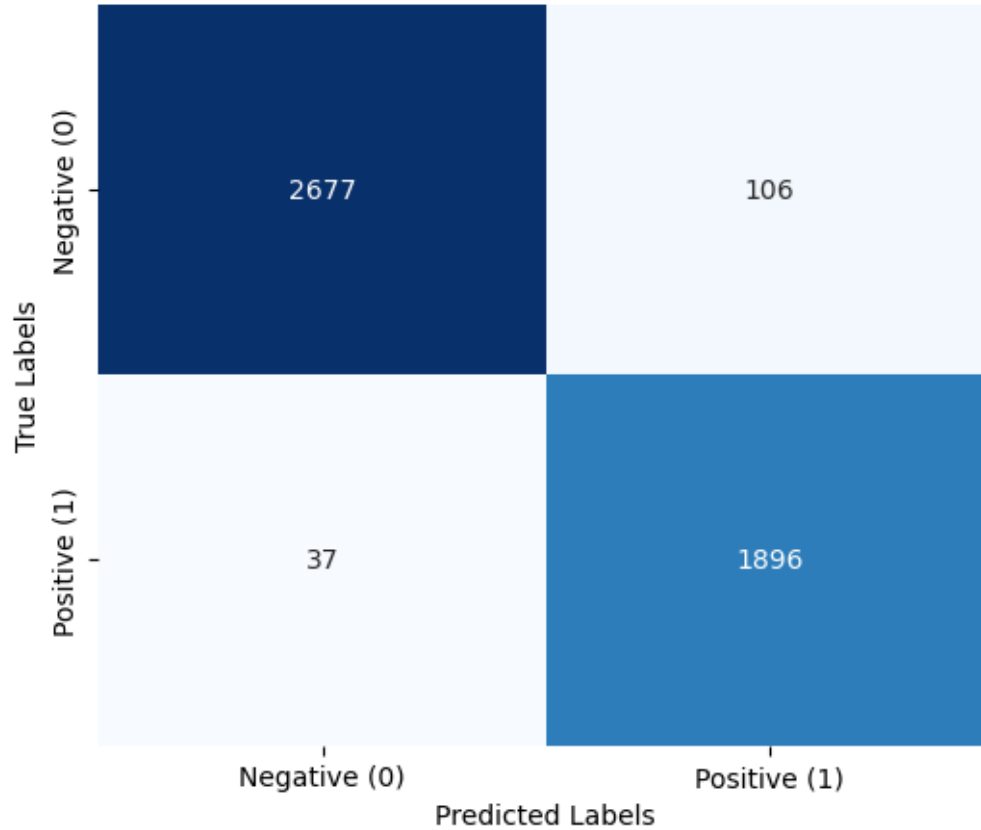


Figure 4.10: Confusion Matrix of MarBERT Model for Binary Sentiment Classification

Comparative evaluation of AraBERT and MarBERT shows that the two transformer-based models have quite competitive results in the Arabic sentiment classification, but each has its strong points in accordance with its linguistic pretraining emphasis. AraBERT, which was trained mostly on the corpora of Modern Standard Arabic (MSA), was a bit more precise in recognizing negative sentiment, which could indicate a higher ability to recognize dissatisfaction written in grammatically structured language frequent in official reviews of banks. On the other hand, pretrained on dialectal and social media text, MarBERT a little higher recall on the positive class, or its sensitivity to informal expressions, colloquial praise and dialect-specific signals of satisfaction. Although both models reached the same overall accuracy of 0.97, AraBERT had slightly fewer misclassifications in negative reviews, meanwhile MarBERT was more competent in detecting positive sentiment nuances. These results indicate that AraBERT can be more useful in SA tasks where the language is more standardized and MarBERT can prove beneficial in the scenarios when it is abundant in dialectal variation, which makes both models complementary to each other based on the linguistic characteristics of the target data.

4.5 Topic Modeling and Semantic Interpretation of User Reviews

In order to supplement sentiment classification and offer more information about the concerns of the users, topic modeling was used to reveal the principal themes as they were reflected in the reviews of Arabic mobile banking. Two different methods were tested, namely the probabilistic Latent Dirichlet Allocation (LDA), an algorithm that recognizes the topics according to the patterns of co-occurrence of words, and a topic modeling approach based on large language models that contextualizes the meaning of semantic coherence and context beyond the frequency of the words. This juxtaposition of these methods makes it possible to measure the extent to which conventional unsupervised models are capable of observing trends in Arabic text in comparison to more sophisticated AI-based tools, which can consider the nuances of language and the dialectal diversity of the data, as well as the relationships between this data and its contextual elements.

4.5.1.1 Evaluation of LDA Topic Modeling

The analysis of the LDA topic model aims at assessing the appropriateness of the underlying corpus and the quality of topics obtained, and especially how the model relates to the semantic structure of Arabic mobile-banking reviews. To check whether the topic modeling data were processed, the preprocessed data used in topic modeling had to be first investigated to address whether it was large enough and lexically diverse to estimate topics through the stable topic modeling.

The summary of the final corpora is presented in Table 4.10, the positive part includes 9,664 documents with about 1,067 unique tokens, whereas the negative part includes 13,914 documents with about 2,196 unique tokens. Such asymmetry is not just due to the greater amount of negative feedback but also to its more lexically varied nature, which is in line with the result that users use a broader breadth of expressions to characterize problems than to characterize generic satisfaction.

Table 4.10: Summary of Preprocessed Corpora Used for LDA Topic Modeling

Dataset	Documents	Unique Tokens (Approx.)
Positive Reviews	9,664	$\approx 1,067$
Negative Reviews	13,914	$\approx 2,196$

Building on this foundation, various LDA models were run using various topic counts (K) to find an optimal complexity of a model to use in each sentiment segment. The coherence score C_v , which is an intrinsic measure of topic interpretability and semantic consistency was calculated at each of the values of K in the range 5-15. Figure 4.11 plots the evolution of coherence scores as a function of topic number for positive and negative reviews, revealing that coherence improves markedly for positive reviews when K approaches 8, 10, and 15, whereas negative reviews achieve their best scores around $K = 9$ and $K = 14$ before declining. These patterns indicate that too few topics lead to overly broad, mixed themes, while excessive topic numbers fragment semantically coherent clusters into noise.

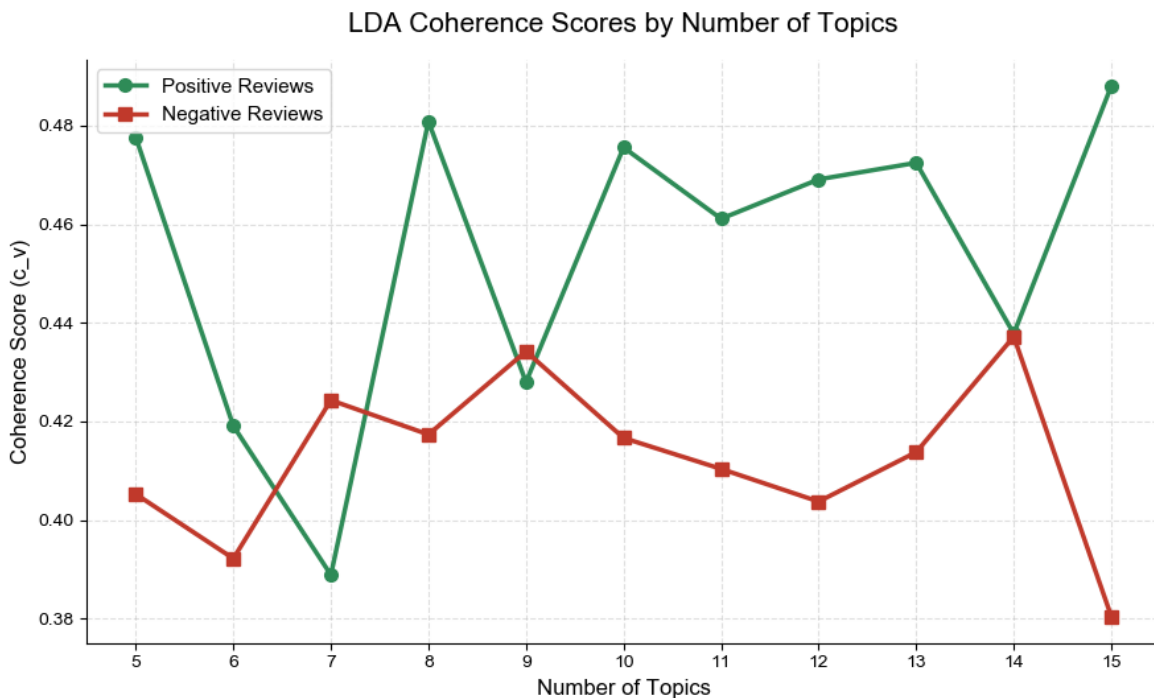


Figure 4.11: Coherence Score Comparison for LDA Topic Models by Number of Topics

On the basis of these curves, and in line with the principle of selecting the simplest model that still provides high semantic clarity, the study adopted 15 topics for positive reviews and 14 topics for negative reviews, as reported in Table 4.11, with approximate coherence scores of 0.49 and 0.44 respectively.

Table 4.11: Selected Number of Topics and Coherence Scores for LDA Models

Sentiment	Selected Topics (K)	Coherence C _v (Approx.)
Positive Reviews	15	≈ 0.49
Negative Reviews	14	≈ 0.44

The increased coherence of the positive subset implies that it is the smaller set of well-defined topics comprising of expressions of satisfaction by the users that are organized around the lower count of subjects (as examples, usability, successful transactions) whereby negative feedback is suitably distributed across a more heterogeneous array of problems (including failures in the logins, crashes and slow performance) which naturally reduces topic compactness. The assessment tasks done confirm that the selected LDA settings represent a fair balance between the complexity of the model and its interpretability and present a solid semantic framework on which the qualitative analysis of the positive and negative user experiences would be based.

Figure 4.12 shows that the bar chart has got most positive reviews concentrated around one large theme with topic ID 2 having about 33.8% of the total corpus. The subject, which is referred to as Excellent App Functionality and User Comfort, is an indication of wide user satisfaction as far as application usability, system smoothness, convenience, and responsiveness are concerned. Keywords such as “ممتاز، رائع، تطبيق، خدمات، مريح، سلس” suggest that users focus extensively on overall ease of use and comfort when interacting with banking apps. In addition to the major theme, some minor but significant ones, including successful digital transactions (Topic 8, 10.9%), a convenient navigation (Topic 9, 9.8%), and service improvements (Topic 7, 9.6%), show that the users appreciate efficient operation and unceasing service improvement. The other items with less than 6 percentage points represent complementary positive experiences like appreciation of customers, secure log in, and reliable banking services. Together, the spread of positive subjects

indicates that satisfaction is more likely to emerge as a result of a consistent combination of functional and experience attributes, in particular, simple usage and fluid financial operations.

Topic Distribution in Positive Reviews

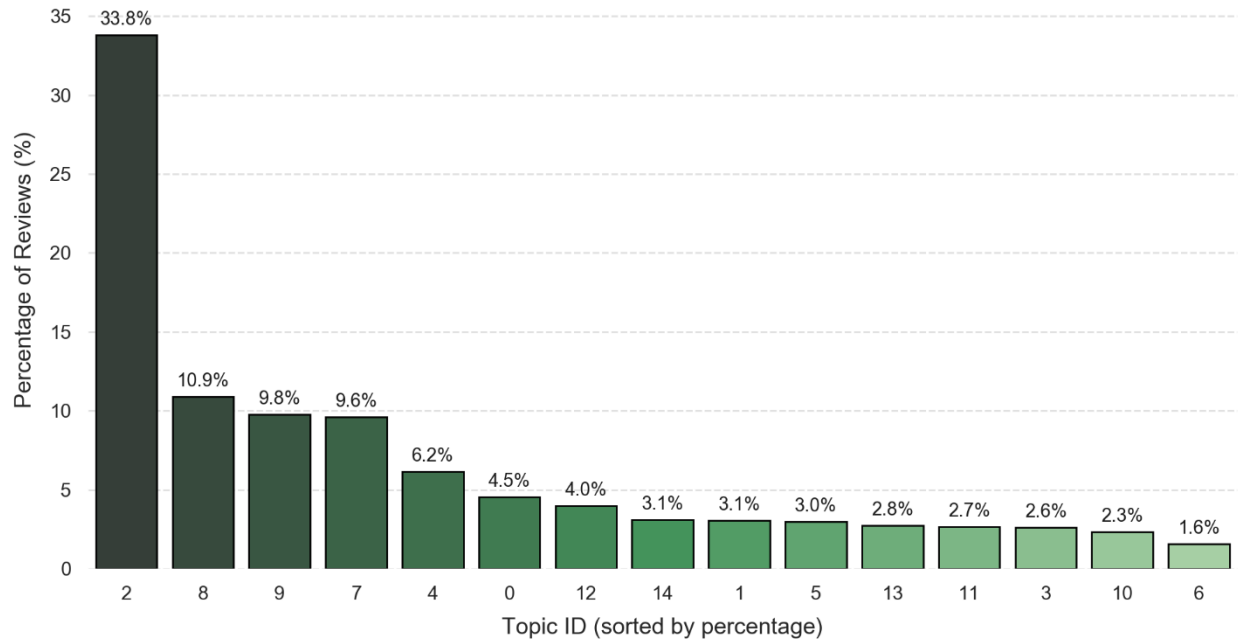


Figure 4.12: Topic Distribution in Positive Reviews Using LDA

Conversely, Figure 4.13 demonstrates a much different trend among negative reviews as dissatisfaction in the case is dominated by only one negative aspect. ID 5, which is the subject of the negative reviews on the issue of the failures to make a login and malfunctioning updates, occupies 94.6 percent of all negative reviews which is the concentration of user frustrations on the issue of accessing the system and malfunctioning of the updates. Frequent terms like “لا يعمل، الدخول، مشكلة، يعلق، التحديث الأخير” reveal systemic failures preventing basic banking activities such as logging into the application. The other issues with less than 2% are indicative of secondary yet significant pain points such as credential errors (Topic 8, 1.6%), transfer difficulties (Topic 6, 1.1%), and complicated procedures that impact the service usability (Topic 13, 0.4%). The representation of the log scale underlines the disproportionate size of Topic 5, indicating that negative feedback is not of general dissatisfaction but of particular, dramatic breakdowns that severely impair the transactional and access to accounts functions.

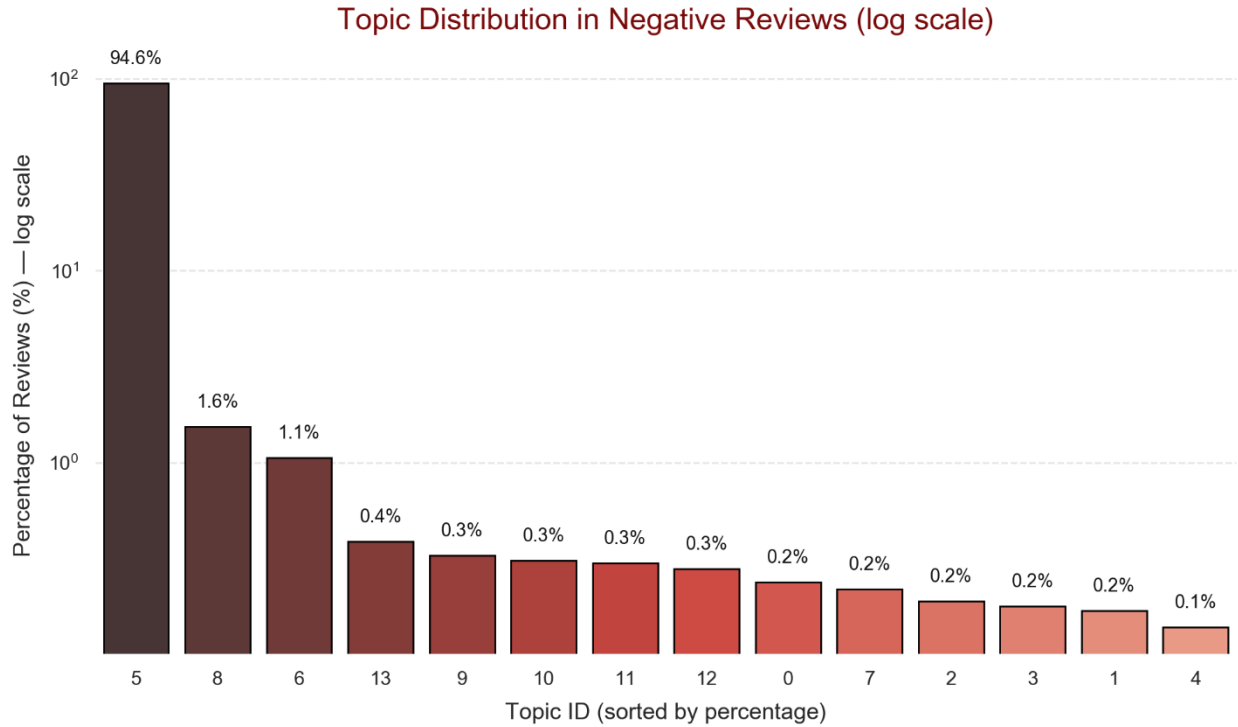


Figure 4.13: Topic Distribution in Negative Reviews Using LDA

This imbalance brings out a key point, whereas on the positive side, one can see general admiration of usability and comfort, negative feedback is mainly caused by failure to perform the basic banking activities. Accordingly, reliability, particularly in the process of making a login and update, becomes a determining element affecting the confidence and attitude of the users to the mobile banking services.

The sentiment-oriented word cloud visualization in figure 4.14 shows the most salient words in both the positive and negative LDA topics which is an intuitive picture of the linguistic patterns that define user perceptions of the mobile banking services. The green word clouds for positive topics (e.g., *Ease of Use*, *General Positive Experience*, and *Customer Service*) reveal a recurrent emphasis on words such as “ممتاز، سريع، رائع، جميل، سهلة، تعامل، خدمة”، which reflect widespread satisfaction with usability, speed, and service interactions. Conversely, the red word clouds for negative topics (e.g., *Technical Issues*, *Login & Access Problems*, and *Slow Performance*) display dominant keywords including “فشل، تعليق، تسجيل، مشكلة، بطيء، تحديث، لا يعمل”، indicating that sentiment shifts sharply when essential functionality breaks down. This visual comparison supports the previously mentioned topic modeling results: a positive experience and well-

performing services are likely to make users satisfied, and the opposite is also possible, as the negative emotion is mostly evoked by accessibility issues and instability.

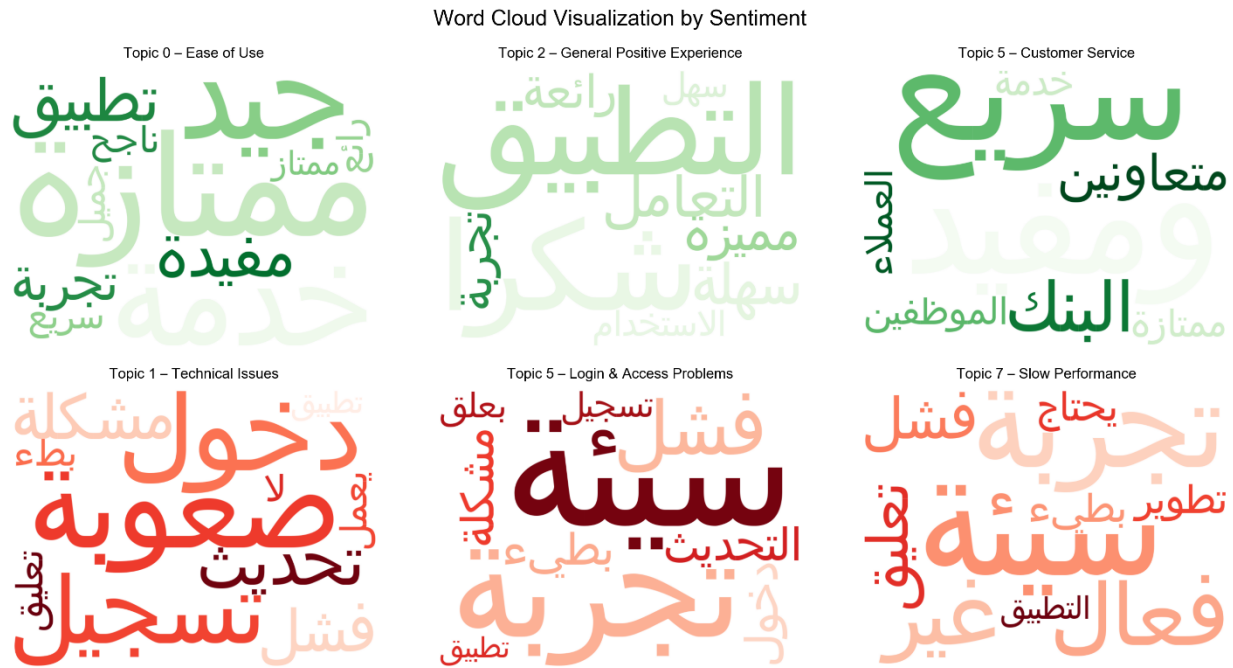


Figure 4.14: Word Cloud Visualization by Sentiment

Figures 4.15 and 4.16 further explain the LDA topic models semantically, by giving the interactive topical visualization of the Latent Dirichlet Allocation (LDAvis) tool to visualize the relative distance of the topics, as well as to show the most salient terms pertinent to each theme. The LDAvis interface to positive bank application reviews is depicted in figure 4.15, the topics being depicted as circles drawn within a two-dimensional semantic space created by multidimensional scaling. The larger circles represent the percentage of the corpus of the respective topic, which proves their prevalence. The fact that there is a gap between the circles in this visualization proves that different thematics are divided, and the positive user feedback is not dispersed randomly but concentrated around consistent constructs like ease of use, operational efficiency, and customer service satisfaction. The bar chart accompanying the map reveals the most relevant terms for the selected dominant topic (Topic ID 2), which includes high-frequency expressions such as “رائع، تجربة، ممتاز، التطبيق، الخدمات، مريح، تجربة”. These terms reinforce the interpretation that users primarily value smooth functionality, intuitive navigation, and reliable performance as core drivers of positive sentiment.

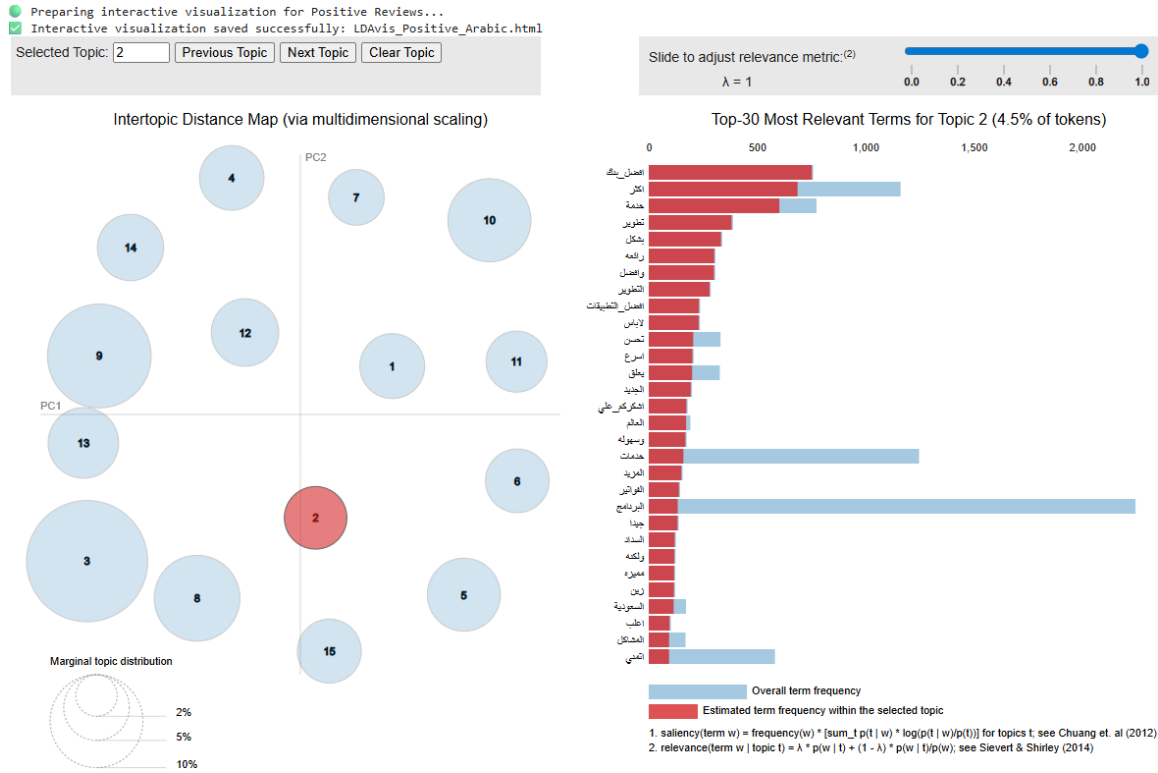


Figure 4.15: LDAvis Interactive Visualization of Positive Topics in Arabic Banking Reviews

Contrarily, Figure 4.16 shows the LDAvis visualization of negative reviews, where the topical dominance is highly biased to a critical pain point, that is, failure to log in and updates malfunctions (Topic ID 5). The large circle is a manifestation of the topic relative to the other topics, and graphically validates that this topic has an enormous proportion in the negative corpus, as has been found previously in the topic distributions. The dense concentration of other smaller circles in the map indicates that complaints that are caused by errors (ex: repeated login attempts, slow transactions, crashes, slow update performance) are semantically related and are based on common disappointments of core application functionality, rather than single problems. The bar chart corroborates this pattern, showing salient negative terms such as “تسجيل، التحديث، فشل، الدخول، يعلق، “مشكلة، لا يعمل”، which reflect recurrent user frustration with authentication failures and non-functional features following updates. The prevalence of complaints based on reliability indicates a crucial point: although users enjoy sophisticated features in cases of its functionality, the slight failure of basic banking features quickly turns into acute dissatisfaction because of the risk involved in financial operations.

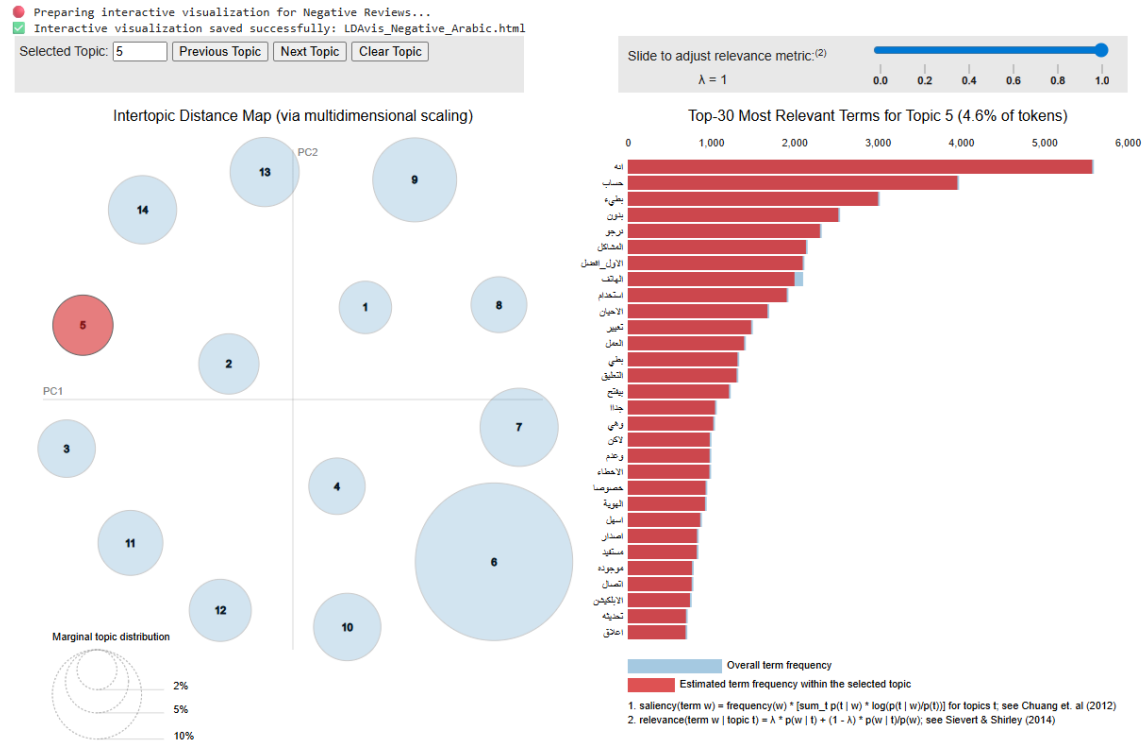


Figure 4.16: LDAvis Interactive Visualization of Negative Topics in Arabic Banking Reviews

The implications of these insights include the preference that banks must optimize system reliability, specifically with regard to updates and user authentication first, before making investments in feature enhancements. The LDAvis tool provides an actionable interpretative and data-driven view of user experience failures and success through visualizing topic prevalence and term relevance simultaneously, rendering the tool useful in enhancing digital banking services.

4.5.1.2 Evaluation of GPT-Based Topic Modeling

In order to supplement the statistical LDA subjects and make it more interpretable, GPT-based topic modeling was used to produce human-readable thematic labels and detailed semantic descriptions of each of the identified topics. In contrast to LDA that only uses co-occurrence patterns to cluster words together, GPT introduces a linguistic and contextual understanding layer that allows more precise topic naming that is consistent with real-life meaning application. This combined methodology maintains methodological rigor since it uses LDA to extract latent themes, whereas interpretation is done only with GPT. It also guarantees the integrity of sentiment boundaries as GPT was not utilized to classify sentiment or modify ground-truth labels. Because of the rich contextual summaries that it offers, it is possible to comprehend the intended meaning

of users who are either pleased or not pleased with the Arabic digital banking apps, which is one of the necessities in the banking experience studies.

The Table 4.12 is a list of the most common GPT-generated topics in the entire corpus of reviews, regardless of sentiment type. The allocation shows a severe thematic imbalance, the negative topics are the most frequent at the highest frequencies. It has the top five topics of Update Issues of 7.09, Performance Issues of 5.57, Login Problems of 5.02, General App Performance of 3.97, and App Not Working of 2.86 that are all system failures that directly impact access, speed, or reliability of the application. These themes emphasize the importance of technical stability as a factor in creating dissatisfaction to users. Contrarily, positive terms like Ease of Use of 2.11% of 0.97% Feature Requests, Convenience, and Customer Service are present in notably lower proportions. The thematic content of positive subjects is less precise and more experience-oriented, and usually indicates comfort, easy interaction, and ability to cope with the task than individual functional accomplishments. This discrepancy indicates that users anticipate banking applications to work perfectly in default, and only get satisfaction when the usage is pleasant, whereas technical breakdowns are highly eliciting feedback.

Table 4.12: Most Frequent GPT Topics

Topic Name (English)	Topic Name (Arabic)	Sentiment	% of All Reviews
Update Issues	مشاكل التحديث	Negative	7.09%
Performance Issues	مشاكل الأداء	Negative	5.57%
Login Issues	مشاكل تسجيل الدخول	Negative	5.02%
App Performance	أداء التطبيق	Negative	3.97%
App Not Working	التطبيق لا يعمل	Negative	2.86%
App Crashes	انهيار التطبيق	Negative	2.79%
App Not Opening	التطبيق لا يفتح	Negative	2.50%
Ease of Use	سهولة الاستخدام	Positive	2.11%

Slow Performance	بطء الأداء	Negative	1.96%
App Freezing	تجمد التطبيق	Negative	1.78%
Frequent Updates	التحديثات المتكررة	Negative	1.35%
Feature Request	طلب إضافة ميزة	Positive	0.97%
Convenience	الراحة والملاءمة	Positive	0.71%
Customer Service	خدمة العملاء	Positive	0.65%
Security	الأمان والحماية	Positive	0.56%
User Experience	تجربة المستخدم	Positive	0.39%
Language Support	دعم اللغة	Positive	0.36%
Service Quality	جودة الخدمة	Positive	0.32%

This observation is further narrowed down by Table 4.13 to only two topics that are shared in both sentiment categories Performance Issues and App Performance. These themes occur in both the satisfied and dissatisfied users, but they are extremely more prevalent in the former category: 5.57 percent negative v/s 0.38 percent positive with Performance Issues and 3.97 percent negative v/s 0.31 percent with App Performance. The contrast is clearly used to show that users think performance improvements are treated as a normal condition, but performance failures have an out of proportion emotional influence and provoke negative feeling and response. Such an asymmetry is consistent with the user-experience theory that has held that negative service experiences have a greater impact on perceptions, loyalty and behavior, especially in high-trust markets like financial services.

Table 4.13: Shared Topics Across Positive and Negative Reviews

Topic (English)	Topic (Arabic)	Positive %	Negative %	Dominant Polarity
Performance Issues	مشاكل الأداء	0.38%	5.57%	Negative

App Performance	أداء التطبيق	0.31%	3.97%	Negative
-----------------	--------------	-------	-------	----------

Figure 4.17 is a graphical representation of the count of positive and negative GPT-interpreted topics, a graphic illustration that is interesting to compare. As the bar chart shows, the user talk is mostly filled with negative items, particularly, those that are related to authentication failures, constant crashes, the inability to open the application, and unstable features post-updates. The positive part, on the other hand, is headed by generic satisfaction descriptors, namely, Ease of Use, Service Quality, Convenience, and Customer Support, which are much less frequent and less intense. This visual presentation supports the idea that technical reliability is the heart of the expectation of users, and positive experiences are appreciated but not strongly expressed. The fact that negative subject matter is predominant (Login Problems, Update Failures and System Performance Issues) has been used to show how sensitive the user is to any form of inconvenience in the banking services where failure to access directly translates into insecurity and loss of confidence.

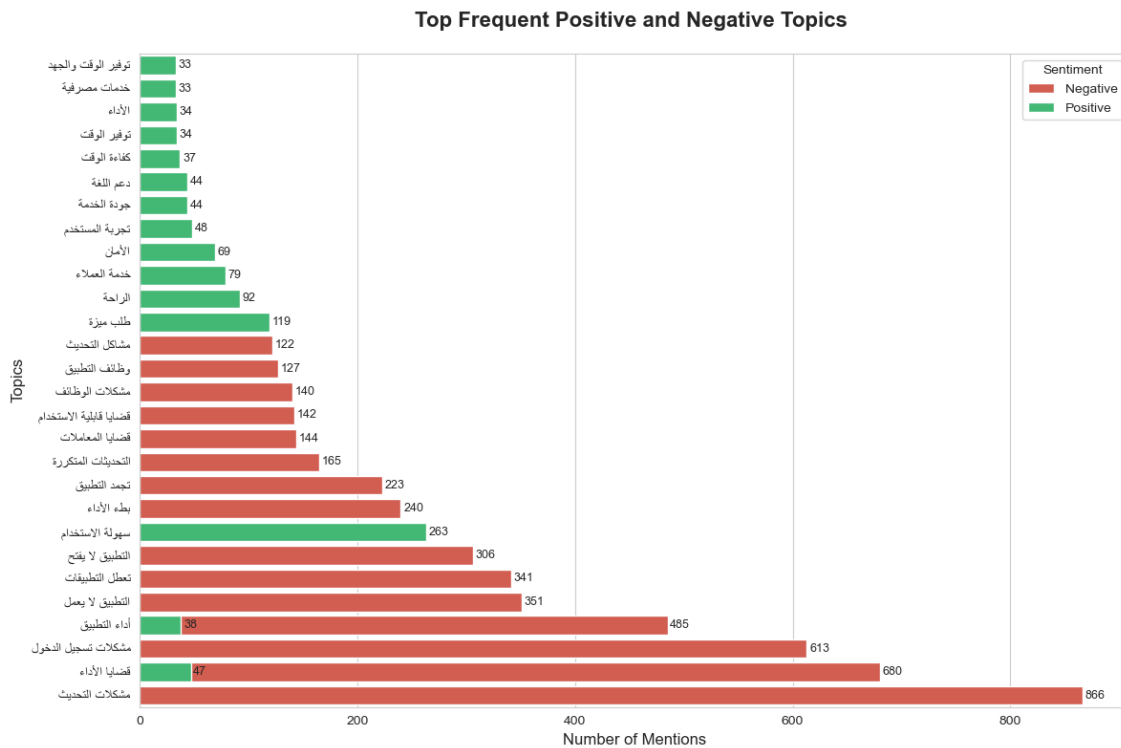


Figure 4.17: Top Frequent Positive and Negative Topics Identified by GPT-Based Topic Modeling

The topic modeling based on GPT indicates that although Arabic mobile banking users like their experience of using the system with ease and efficiency, almost all the dissatisfaction is determined by technical instability and non-reliability. These findings suggest that strategic change in authentication stability, system speed and post-update quality control would be the most significant to overall satisfaction. Hence, technical strength should be the focus instead of increasing features or making the applications look better since the user feeling in banking apps is trust oriented and reliability oriented.

4.5.1.3 Comparative Analysis of LDA and GPT-Based Topic Modeling

To reinforce the thematic analysis of Arabic mobile banking reviews, a combination of two complementary topic modeling tools the probabilistic LDA and semantics-driven GPT-based topic interpretation is employed. Despite the fact that both methods were used on the same data, these approaches have a significant difference in view of their analytical paradigms and linguistic assumptions. LDA is based on the statistical inference to reveal latent word co-occurrence patterns in text corpora and needs extensive pre-processing and domain interpretation to create meaningful topic names. GPT, by contrast, is a Large Language Model (LLM) which uses contextual reasoning and semantic abstraction directly to natural language input, and generates human readable topic assignments without directly modeling word distributions. The methodological rigor is achieved by the comparative use of both methodologies: LDA shows the latent statistical form of the opinions of users, and GPT gives the interpretation the descriptive labeling in accordance with linguistic manifestations of users. The approximate comparison of the two topic modeling methods has been arranged in Table 4.14 and it highlights the complementary aspects of probabilistic and language model driven methods of analyzing user-generated banking reviews.

LDA in this study was simply trained on positive and negative subsets because of the imbalance in polarity and the different lexical signatures that are linked to satisfaction and dissatisfaction with Arabic online banking. To establish the best number of topics, topic coherence was employed, which produced 15 themes in positive feedback and 14 in negative reviews, and thus, providing semantic understanding and minimizing repetitive groupings. GPT, however, did not need topic number estimation, it produced natural-language topic names of single reviews and it enabled the aggregation of prevailing themes in frequency distribution. Although LDA found granular co-occurrence groupings, GPT labels were shorter and more thematically consistent, more

comparable to how users characterize real-world problems, like Update Issues, Slow Performance or Ease of Use.

Table 4.14: Comparison Between LDA and GPT-Based Topic Modeling

Aspect	LDA Topic Modeling	GPT-Based Topic Modeling
Model Type	Probabilistic generative model (based on word co-occurrence and Dirichlet priors).	Large Language Model generating semantic topic labels and textual summaries.
Input Representation	Requires tokenization, stopword removal, bigrams; trained separately for positive and negative corpora.	Operates directly on full-length Arabic text without extensive preprocessing.
Number of Topics	Determined through coherence scoring (15 positive, 14 negative).	No fixed topic number; themes emerge from aggregated GPT topic assignments.
Output Format	Lists of high-probability keywords requiring manual interpretation.	Human-readable topic names and descriptive summaries aligned with user intent.
Interpretability	Indirect; requires expert labeling to make topics meaningful.	Immediate; produces clear semantic labels with minimal post-processing.
Granularity of Themes	Captures statistical co-occurrence structures; may blend sub-themes.	Produces precise, concept-focused themes reflecting natural language usage.
Sentiment Handling	Must train separate models to avoid polarity mixing.	Automatically captures sentiment alongside topic assignment.

Strengths	Data-driven, reproducible, interpretable via coherence metrics.	Highly interpretable, context-aware, dialect-sensitive, and human-aligned.
Limitations	Requires manual labeling; may capture overly broad or mixed topics.	Dependent on prompt quality; not a statistical representation of latent structure.

The conceptual match between models showed significant point of convergence in the case of core issues like failure to log in, application crash, malfunction to update and reliability of the transaction and comfort in usability. The co-occurrence of LDA and GPT furnishes the significance of these themes as the powerful factors of user experience in Arabic mobile banking apps. Nonetheless, they were not identical in the thematic granularity: LDA occasionally combined praise and usability in a similar topic, whereas GPT split them into different sentiment-sensitive categories. This deviation indicates the statistical categorization of lexical co-occurrences of LDA as compared to the semantic processing of lexical interpretive meaning of GPT. By incorporating both methods, therefore, one will end up with a deeper analytical model: LDA will gain more empirical foundations and GPT will gain more interpretability and explanatory transparency. Such methodological complementarity is necessary in this sphere where language expressions usually involve an amalgamation of emotion, derision, or dialectal fluctuations, to achieve a better insight into the expectation and dissatisfaction of the user.

4.6 Conclusion

The empirical study developed in the course of this chapter shows that the combination of the traditional ML, DL, and large language modeling methods is effective in terms of Arabic sentiment and topic analysis in mobile banking applications. The relative analysis indicates that albeit classical algorithms have the advantage of appropriate data balancing techniques, DL models especially the LSTM networks can be better in their performance due to the presence of the contextual and sequential dependencies in the Arabic text. The models based on LLM, in particular, AraBERT and MarBERT, advance the field of sentiment prediction, as they utilize the pre-trained linguistic knowledge that is related to Arabic morphology and dialects. Along with sentiment

classification, topic analysis findings indicate that the greatest portion of user satisfaction is caused by usability and easy digital transactions, whereas dissatisfaction is caused by technical failures, difficult logins, and unreliable updates. Notably, the overlap of LDA and GPT topic interpretation makes the identified themes more valid and emphasizes the important system qualities that users appreciate. The results verify that sentiment-based and semantically informed analytics have the potential to deliver practical information on improving customer-facing financial technology, which highlights the importance of information-based intelligence in digital banking.

5. Conclusion and Future Work

5.1 Conclusion

Mobile banking application has become a primary factor of financial engagement and service provision due to the exponential growth in digitalization within the Arab world. This has made the perception of the Arabic-speaking customers not only a technological issue, but a necessary process of gaining customer faith, better service delivery and competitiveness of banks operations.

The current research examined the SA of the banking industry within the context of a linguistically challenging setting, whereby users describe their experiences with the use of a blend of both the modern standard Arabic and various dialects. These terms are mainly characterized by informal structures, cultural idioms and cues that can only be deciphered by their context and the traditional methods of analysis can do nothing.

Thus, the examination of opinionated customer-generated reviews of mobile banking application can be strategic in terms of understanding customer behavior and emotional reactions and expectations, and it will also give banks the chance to make digital services better and raise customer satisfaction in Arabic-speaking markets.

SA in the Arabic banking is still underdeveloped as it is a topic that is vital but is adversely affected by the language barrier of the Arabic language. These issues are morphological richness, dialectal diversity, non-standardized spelling and the lack of well-organized linguistic corpora, particularly in the financial spheres. Most of the available research uses small samples, generic lexicons, or inconsistent and biased annotation practices, which result in untrustworthy assessments and inaccurate generalization. The inability of the classical ML models to represent context, and the reliance of DL models on large well-balanced datasets also adds a problem to the advances in Arabic SA.

This thesis suggested a sentiment and topic analysis system, which was designed to process noisy and dialect-containing Arabic reviews and derive sentiment polarity and thematic issues with a greater level of accuracy and interpretability. The model integrates high-level preprocessing based on Arabic, multi-source annotation strategies, balanced sampling, as well as, and the use of ML, DL, and large language models, and then a hybrid topic modeling model that integrates probabilistic and semantic models.

The proposed research design is presented in a systematic sequence of steps, the first of which is the massive data mining of Google Play, which includes the systematic text normalization aimed at preserving the content with the sentiment and minimizing linguistic noise. This is followed by a novel hybrid annotation process that incorporates manual labeling, rating-based labeling, and GPT-assisted labeling in order to enhance consistency and decrease subjectivity. Due to the common use of the Arabic datasets of class imbalance, several balancing methods were used independently with class weights, SMOTE and random oversampling to enable fair comparative analysis across the algorithms. Then, TF-IDF feature extraction allowed ML methods, including SVM, Logistic regression, Naive Bayes and random forest to classify sentiments and Long short term memory (LSTM) modeled sequential relationships in free-form reviews. And lastly, the state-of-the-art Arabic large language models were introduced: AraBERT and MARBERT which made use of contextual insights and dialectal coverage, providing a clear-cut advantage over alternative methods.

These experimental findings proved the effectiveness of deep contextual language models as compared to the traditional ones. Although ML models could be competitive using balanced data, they could not interpret complex language structures making their performance limited. The ability to learn order of words dependencies enhanced LSTM models, particularly in cases where SMOTE and oversampling techniques were used to support the models.

However, the biggest improvements were made on the large language models, most notably the MARBERT, that surpassed all other classifiers because they managed the dialectal expressions, informal writings style, and semantically rich context that occur in Arabic banking reviews. These results indicate that classification of sentiment in Arabic financial sectors needs models that can respond to contextual change and not just statistical term frequency or sequential dependencies.

Besides sentiment classification, this thesis also developed its analysis and uncovered the semantic themes that resulted in user satisfaction and dissatisfaction with the use of topic modeling. LDA permitted an approximate distribution of repeated subjects with positive and negative sentiment subsets, and GPT simplified them into labels that can be read and understood by humans, as well as labels that are semantically consistent. The negative issues prevailed in the corpus, and the common themes were crashes of apps, failure to log in, slow functionality, and malfunctioning after updating. However, there were fewer positive themes and they concerned ease of use,

convenience, and customer service which indicated that the user would be more responsive to technical breakdowns rather than positive experiences. Such imbalance reflects the behavioral pattern of digital service industries: users have become used to successful functionality, but are extremely judgmental when a system not meets expectations. Indicating that the increase in performance, stability, authentication procedures and update mechanisms would lead to disproportional returns in customer satisfaction.

This thesis provided an analytical framework of sentiment and topic detection in Arabic banking reviews that is comprehensive and scalable and overcomes the linguistic complexity and data limitations that have impeded research in this area in the past. This work combines ML, DL, and highly developed LLMs, probabilistic and semantic topic extraction to make a contribution to Arabic Natural Language Processing at the theoretical and practical levels. Future studies can expand on these results and aim to apply domain-adaptive pretraining on Arabic financial language models, make global banking markets more multilingual, and implement real-time sentiment detection systems with the potential to make strategic decisions in the digital banking landscape of the MENA region.

The proposed framework exhibits strong cross-domain generalizability, as its core pipeline addresses linguistic challenges—dialectal variation, sarcasm, code-switching, and informal register—that are inherent to Arabic user-generated content rather than specific to the banking sector. Consequently, the preprocessing, annotation, classification, and topic modeling components can be transferred to other service industries, including telecommunications, e-commerce, healthcare, and hospitality, where customer reviews exhibit structurally identical NLP challenges.

Critically, the modular architecture facilitates this adaptation with minimal effort: only the domain-specific training data requires replacement, while the analytical infrastructure remains intact. This positions the system not as a single-use banking tool, but as a reusable Arabic sentiment analysis foundation applicable across the broader MENA digital services ecosystem. Future research should empirically validate this transferability through cross-domain comparative experiments.

It is important to note several limitations of this work: the manual annotation was performed by a single annotator (making the labels reference labels rather than absolute ground truth), GPT-4o outputs may vary across API calls (introducing reproducibility constraints), and the dataset is

limited to Google Play reviews which may not capture all customer sentiment channels. Future research should address these limitations through multi-annotator frameworks, deterministic labeling alternatives, and multi-platform data collection.

5.2 Future Work

Although this study creates a solid and generalizable framework of Arabic sentiment and topic analysis within the framework of mobile banking application, there are a number of promising directions that can add to its scientific and practical value.

To begin with, future research might increase the data set beyond the reviews of app stores to multi-source and multi-channel insight of user experience by adding real-time social media feedback, chatbot interactions, and customer support logs to it. The other significant direction of development is the domain-specific Arabic language model fine-tuning with financial terms, dialectal differences, as well as with Arabic language behavioral patterns specific to the digital banking customer. The strategies of continuous learning might help such models to become more dynamic and meet any new banking features, the development of user expectations, and language changes in the usage of dialects.

Furthermore, the real-time sentiment monitoring systems might also be introduced in the form of live dashboards to banks to convert the outputs of the analytical tools into the actionable insights that would be used to make the decisions, prevent risks, detect fraud, or retain clients.

Multimodal SA could also be addressed in future studies by combining the text-based information with metadata of user behavioral patterns either regarding the frequency of logins, duration of the session, failure rates of transactions, or the device attributes that could yield a more insightful view of the nature of emotion-based usability problems. Last but not least, the existing method of cross-lingual and transfer learning, which is the extension of current approach to a multilingual environment, may allow conducting comparative studies of regional and international markets, in which the Arabic digital banking analytics would be placed in a global context. Combined, these directions will provide a productive way of customer-centric solutions, digital financial services, and innovation in the research of Arabic fintech data.

References

- Al-Qudah, A. A., Al-Okaily, M., Alqudah, G., & Ghazlat, A. (2024). Mobile payment adoption in the time of the COVID-19 pandemic. *Electronic Commerce Research*, 24(1), 427-451.
- Alkhwaldi, A. F., Alharasis, E. E., Shehadeh, M., Abu-AlSondos, I. A., Oudat, M. S., & Bani Atta, A. A. (2022). Towards an understanding of FinTech users' adoption: Intention and e-loyalty post-COVID-19 from a developing country perspective. *Sustainability*, 14(19), 12616.
- Statista, "Digital population worldwide as of january 2024," <https://www.statista.com/statistics/617136/digital-populationworldwide/>, Feb. 2024, accessed on 9 august 2025.
- Persia, F., & D'Auria, D. (2017, August). A survey of online social networks: challenges and opportunities. In *2017 IEEE international conference on information reuse and integration (IRI)* (pp. 614-620). IEEE.
- Murthy, G. S. N., Allu, S. R., Andhavarapu, B., Bagadi, M., & Belusonti, M. (2020). Text based sentiment analysis using LSTM. *Int. J. Eng. Res. Tech. Res*, 9(05), 32-41.
- Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020). Sentiment analysis using gated recurrent neural networks. *SN Computer Science*, 1(2), 74.
- Michailidis, P. D. (2024). A Comparative Study of Sentiment Classification Models for Greek Reviews. *Big Data and Cognitive Computing*, 8(9), 107.
- Sharma, H. D., & Goyal, P. (2023). An analysis of sentiment: methods, applications, and challenges. *Engineering Proceedings*, 59(1), 68.
- De Keyser, A., Lemon, K. N., Klaus, P., & Keiningham, T. L. (2015). A framework for understanding and managing the customer experience. *Marketing Science Institute working paper series*, 85(1), 15-121.
- Lin, H., Gursoy, D., & Zhang, M. (2020). Impact of customer-to-customer interactions on overall service experience: A social servicescape perspective. *International Journal of Hospitality Management*, 87, 102376.
- Kumar, P., Mokha, A. K., & Pattnaik, S. C. (2021). Electronic customer relationship management (E-CRM), customer experience and customer satisfaction: evidence from the banking industry. *Benchmarking: An International Journal*, 29(2), 551-572.
- Mokha, A. K., & Kumar, P. (2022). Examining the interconnections between E-CRM, customer experience, customer satisfaction and customer loyalty: A mediation approach. *Journal of Electronic Commerce in Organizations (JECO)*, 20(1), 1-21.
- Garzaro, D. M., Varotto, L. F., & de Carvalho Pedro, S. (2020). Internet and mobile banking: the role of engagement and experience on satisfaction and loyalty. *International Journal of Bank Marketing*, 39(1), 1-23.
- Garg, R., Rahman, Z., & Qureshi, M. N. (2014). Measuring customer experience in banks: scale development and validation. *Journal of Modelling in Management*, 9(1), 87-117.

- Chauhan, S., Akhtar, A., & Gupta, A. (2022). Customer experience in digital banking: A review and future research directions. *International Journal of Quality and Service Sciences*, 14(2), 311-348.
- Ramaswamy, S., Khande, R., Patil, Y. S., & Kalkar, P. (2021). Adapting digital banking services during covid-19 pandemic: a study on customer experiences in selected district of maharashtra state, India. *Journal of Electrical Engineering and Technology (IJEET)*, 12(5), 8-17.
- Ashrafpour, N., Niky Esfahlan, H., Aali, S., & Taghizadeh, H. (2022). The prerequisites and consequences of customers' online experience regarding the moderating role of brand congruity: evidence from an Iranian bank. *Journal of Islamic Marketing*, 13(10), 2144-2172.
- Shaikh, A. A., Alharthi, M. D., & Alamoudi, H. O. (2020). Examining key drivers of consumer experience with (non-financial) digital services—An exploratory study. *Journal of Retailing and Consumer Services*, 55, 102073.
- Moliner-Tena, M. A., Monferrer-Tirado, D., & Estrada-Guillén, M. (2019). Customer engagement, non-transactional behaviors and experience in services: A study in the bank sector. *International Journal of Bank Marketing*, 37(3), 730-754.
- Mbama, C. I., & Ezepue, P. O. (2018). Digital banking, customer experience and bank financial performance: UK customers' perceptions. *International journal of bank marketing*, 36(2), 230-255.
- Szmigin, I., & Piacentini, M. (2018). *Consumer behaviour*. Oxford University Press.
- Oskamp, S., & Schultz, P. W. (2005). *Attitudes and opinions*. Psychology Press.
- Dauda, S. Y., & Lee, J. (2015). Technology adoption: A conjoint analysis of consumers' preference on future online banking services. *Information Systems*, 53, 1-15.
- Mwanajimba, Z. N. (2019). *Attitude toward self-service technologies among retail customers of commercial banks in Tanzania: The case of National Micro-finance Bank* (Doctoral dissertation, The University of Dodoma).
- Sakhaei, S. F., Afshari, A. J., & Esmaili, E. (2014). The impact of service quality on customer satisfaction in Internet banking. *Journal of mathematics and computer science*, 9(1), 33-40.
- Shrestha, D., Wenan, T., Rajkarnikar, N., & Jeong, S. R. (2020). Consumers attitude towards Internet banking services in an underdeveloped country: A case of Pokhara, Nepal. *Journal of Internet Computing and Services*, 21(5), 75-85.
- Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, 6, 89-96.
- Srivastava, S., Singh, M. K., & Singh, Y. N. (2021). Social media analytics: current trends and future prospects. In *Communication and Intelligent Systems: Proceedings of ICCIS 2020* (pp. 1005-1016). Springer Singapore.

- Lauer, D. (2021). Facebook's ethical failures are not accidental; they are part of the business model. *AI and Ethics*, 1(4), 395-403.
- Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65).
- Neri, F., Aliprandi, C., Capeci, F., & Cuadros, M. (2012, August). Sentiment analysis on social media. In *2012 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 919-926). IEEE.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., ...Zhu, R.-J. (2023). RWKV: Reinventing RNNs for the Transformer Era. arXiv, 2305.13048. Retrieved from <https://arxiv.org/abs/2305.13048v2>
- Şahin, G. (2020). Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics. *Computational Linguistics*, 46(2), 511–514. doi: 10.1162/coli_r_00381
- Rai, A., & Borah, S. (2020). Study of Various Methods for Tokenization. *Applications of Internet of Things*. Springer. doi: 10.1007/978-981-15-6198-6_18
- Roy, A. (2021). Recent Trends in Named Entity Recognition (NER). arXiv, 2101.11420. Retrieved from <https://arxiv.org/abs/2101.11420v1>
- K. Jiang and X. Lu, "Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review," 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), Chongqing City, China, 2020, pp. 210-214, doi: 10.1109/IICSPI51290.2020.9332458.
- Khondaker, M. T. I., Waheed, A., & Abdul-Mageed, M. (2023, December). Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 220-247).
- Al-Hagree, S., & Al-Gaphari, G. (2022, October). Arabic Sentiment Analysis Based Machine Learning for Measuring User Satisfaction with Banking Services' Mobile Applications: Comparative Study. In *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)* (pp. 1-4). IEEE.
- Nafea, A. A., Muayad, M. S., Majeed, R. R., Ali, A., Bashaddadh, O. M., Khalaf, M. A., ... & Steiti, A. (2024). A brief review on preprocessing text in Arabic language dataset: Techniques and challenges. *Babylonian Journal of Artificial Intelligence*, 2024, 46-53.
- Yang, X., Zhao, H., Xu, W., Qi, Y., Lu, J., Phung, D., & Du, L. (2025, July). Neural topic modeling with large language models in the loop. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1377-1401).

- A. Semary, N., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. *PLoS One*, 19(2), e0294968.
- Maree, M., Eleyat, M., Rabayah, S., & Belkhatir, M. (2023). A hybrid composite features based sentence level sentiment analyzer.
- Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.
- Rizvi, S. A. M., Tasneem, F. B., Ahmad, R., Khan, M. U., Jamshed, N., & Ahmad, T. (2024, November). Exploring the Efficacy of Large Language Models in Topic Modelling: A Comparative Analysis. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)* (pp. 1-6). IEEE.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).
- Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior therapy*, 51(5), 675-687.
- Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to machine learning for physicists. *Physics reports*, 810, 1-124.
- Razali, N. A. M., Malizan, N. A., Hasbullah, N. A., Wook, M., Zainuddin, N. M., Ishak, K. K., ... & Sukardi, S. (2021). Opinion mining for national security: techniques, domain applications, challenges and research opportunities. *Journal of big data*, 8, 1-46.
- Rani, M., & Singh, J. (2021). A primer on opinion mining: the growing research area. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1* (pp. 1059-1076). Springer Singapore.
- Khan, A., Younis, U., Kundi, A. S., Asghar, M. Z., Ullah, I., Aslam, N., & Ahmed, I. (2020). Sentiment classification of user reviews using supervised learning techniques with comparative opinion mining perspective. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 2 1* (pp. 23-29). Springer International Publishing.
- Dhanalakshmi, V., Bino, D., & Saravanan, A. M. (2016, March). Opinion mining from student feedback data using supervised learning algorithms. In *2016 3rd MEC international conference on big data and smart city (ICBDSC)* (pp. 1-5). IEEE.
- Pamina, J., Raja, B., SathyaBama, S., Sruthi, M. S., & VJ, A. (2019). An effective classifier for predicting churn in telecommunication. *Jour of Adv Research in Dynamical & Control Systems*, 11.
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, 67, 93-104.

- Christiansen, S. D. (2021). *Ischemic stroke thrombus characterization through quantitative magnetic resonance imaging* (Doctoral dissertation, The University of Western Ontario (Canada)).
- Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, 18(60), 1-8.
- Rouder, J. N., & Morey, R. D. (2018). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*.
- Aggarwal, C. C. (2018). *Neural networks and deep learning* (Vol. 10, No. 978, p. 3). Cham: springer.
- Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, 18(10), 1099-1104.
- Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support vector machines for classification. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, 39-66.
- Gan, S., Shao, S., Chen, L., Yu, L., & Jiang, L. (2021). Adapting hidden naive Bayes for text classification. *Mathematics*, 9(19), 2378.
- Blanco, V., Japón, A., & Puerto, J. (2020). Optimal arrangements of hyperplanes for SVM-based multiclass classification. *Advances in Data Analysis and Classification*, 14(1), 175-199.
- Usmani, S., Saboor, A., Haris, M., Khan, M. A., & Park, H. (2021). Latest research trends in fall detection and prevention using machine learning: A systematic review. *Sensors*, 21(15), 5134.
- Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2), 74-79.
- Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.
- Graves, A., & Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.
- Adytia, D., Saepudin, D., Tarwidi, D., Pudjaprasetya, S. R., Husrin, S., Sopaheluwakan, A., & Prasetya, G. (2023). Modelling of Deep Learning-Based Downscaling for Wave Forecasting in Coastal Area. *Water*, 15(1), 204.
- Xiao, T., & Zhu, J. (2023). Introduction to Transformers: an NLP Perspective. *arXiv preprint arXiv:2311.17633*.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842-866.
- Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, I.

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Prateek Joshi. How do Transformers Work in NLP? A Guide to the Latest State-of-the-Art Models. url: <https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/> (cit. on p. 18).
- Tripathy, J. K., Chakkaravarthy, S. S., Satapathy, S. C., Sahoo, M., & Vaidehi, V. (2022). ALBERT-based fine-tuning model for cyberbullying analysis. *Multimedia Systems*, 28(6), 1941-1949.
- Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- Abdul-Mageed, M., & Elmadany, A. (2021, August). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7088-7105).
- El Moubtahij, H., Abdelali, H., & Tazi, E. B. (2022). AraBERT transformer model for Arabic comments and reviews analysis. *IAES International Journal of Artificial Intelligence*, 11(1), 379-387.
- Husain, F., & Uzuner, O. (2021, April). Leveraging offensive language for sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 364-369).
- Abdul-Mageed, M., Zhang, C., Elmadany, A., & Ungar, L. (2020). Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.
- Mahmood, T., Naseem, S., Ashraf, R., Asif, M., Umair, M., & Shah, M. (2023). Recognizing factors effecting the use of mobile banking apps through sentiment and thematic analysis on user reviews. *Neural Computing and Applications*, 35(27), 19885-19897.
- Poornima, A., & Priya, K. S. (2020, March). A comparative sentiment analysis of sentence embedding using machine learning techniques. In *2020 6th international conference on advanced computing and communication systems (ICACCS)* (pp. 493-496). IEEE.
- Hawladar, M., Ghosh, A., Raad, Z. K., Chowdhury, W. A., Shehan, M. S. H., & Ashraf, F. B. (2021, September). Amazon product reviews: Sentiment analysis using supervised learning algorithms. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)* (pp. 1-6). IEEE.
- Patrick, H. A., Gujjar, P., Sharief, M. H., & Mukherjee, U. (2023, February). Sentiment Analysis Perspective using Supervised Machine Learning Method. In *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.
- Maree, M., Eleyat, M., & Mesqali, E. Optimizing Machine Learning-based Sentiment Analysis Accuracy in Bilingual Sentences via Preprocessing Techniques. *The International Arab Journal of Information Technology*, 21(2), 257-270.

- Bhowmik, S., Sadik, R., Akanda, W., & Pavel, J. R. (2024). Sentiment analysis with hotel customer reviews using FNet. *Bulletin of Electrical Engineering and Informatics*, 13(2), 1298-1306.
- Tan, K. L., Lee, C. P., Lim, K. M., & Anbananthen, K. S. M. (2022). Sentiment analysis with ensemble hybrid deep learning model. *IEEE Access*, 10, 103694-103704.
- Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE access*, 8, 23522-23530.
- Singh, C., Imam, T., Wibowo, S., & Grandhi, S. (2022). A deep learning approach for sentiment analysis of COVID-19 reviews. *Applied Sciences*, 12(8), 3709.
- Kaur, G., & Sharma, A. (2023). A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of big data*, 10(1), 5.
- Atandoh, P., Zhang, F., Adu-Gyamfi, D., Atandoh, P. H., & Nuhoho, R. E. (2023). Integrated deep learning paradigm for document-based sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, 35(7), 101578.
- Paulraj, D., Ezhumalai, P., & Prakash, M. (2024). A Deep Learning Modified Neural Network (DLMNN) based proficient sentiment analysis technique on Twitter data. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(3), 415-434.
- Husain, F., Alostad, H., & Omar, H. (2024). Bridging the Kuwaiti Dialect Gap in Natural Language Processing. *IEEE Access*.
- Alsemaree, O., Alam, A. S., Gill, S., & Uhlig, S. (2024). Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions. *Heliyon*.
- Omar, A., & Abd El-Hafeez, T. (2023). Quantum computing and machine learning for Arabic language sentiment classification in social media. *Scientific Reports*, 13(1), 17305.
- Khairy, M., Mahmoud, T. M., Omar, A., & Abd El-Hafeez, T. (2023). Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection. *Language Resources and Evaluation*, 1-18.
- Bolbol, N. K., & Maghari, A. Y. (2020, December). Sentiment analysis of arabic tweets using supervised machine learning. In *2020 International Conference on Promising Electronic Technologies (ICPET)* (pp. 89-93). IEEE.
- Abo, M. E. M., Idris, N., Mahmud, R., Qazi, A., Hashem, I. A. T., Maitama, J. Z., ... & Yang, S. (2021). A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection. *Sustainability*, 13(18), 10018.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Alyafeai, Z., Al-shaibani, M. S., Ghaleb, M., & Ahmad, I. (2023). Evaluating various tokenizers for arabic text classification. *Neural Processing Letters*, 55(3), 2911-2933.
- Abudalfa, S. I., Abdu, F. J., & Alowaifeer, M. M. (2024). Arabic Text Formality Modification: A Review and Future Research Directions. *IEEE Access*.

Everington, J. (2024, August 22). The Banker's Top 100 Arab Banks 2024. The Banker.
<https://www.thebanker.com/content/be39b95e-57a6-5ef9-9916-9ac6dafda157>

الملخص

أدى التوسع المتسارع في الخدمات المصرفية الرقمية إلى تعزيز دور تطبيقات البنوك عبر الهاتف المحمول بوصفها قناة رئيسية للتفاعل المالي وتقديم الخدمات في العالم العربي. وتهدف هذه الدراسة إلى تحليل مشاعر المستخدمين الناطقين بالعربية تجاه تطبيقات الخدمات المصرفية، إلى جانب تحديد الموضوعات الرئيسية التي تؤثر في مستويات الرضا وعدم الرضا، بما يساهم في دعم تطوير الخدمات المصرفية الرقمية واتخاذ القرارات التحسينية.

أُجريت الدراسة خلال عام 2025 اعتمادًا على مراجعات المستخدمين المنشورة في متجر Google Play لتطبيقات الخدمات المصرفية التابعة لأفضل 100 بنك عربي وفق تصنيف The Banker لعام 2024. تم جمع البيانات باستخدام أداة استخراج بيانات مطوّرة بلغة Python، وأسفر ذلك عن أكثر من 361,000 مراجعة. وقد جرى ترشيح البيانات للاقتصار على المحتوى العربي فقط، مع الحفاظ على التوزيع الجغرافي الطبيعي للمراجعات في منطقة الشرق الأوسط، بما يعكس أنماط الاستخدام الواقعية وتنوع اللهجات. وتمثل هذه المراجعات مجتمع الدراسة، في حين استُخدمت عينة موسومة منها لأغراض التحليل والنمذجة.

اعتمدت الدراسة منهجية تحليل لغوي مخصصة للغة العربية شملت التطبيع اللغوي، ومعالجة الاختلافات اللهجية، والحفاظ على أدوات النفي، وبناء ثنائيات لفظية ذات دلالة. كما جرى اختبار ومقارنة عدة منهجيات لتوسيم البيانات، شملت التوسيم اليدوي، والاستدلال على المشاعر اعتمادًا على تقييمات المستخدمين، والتصنيف بمساعدة نماذج GPT، حيث تم اعتماد التوسيم اليدوي كمرجعية نهائية نظرًا لما أظهره من مستوى أعلى من الاتساق والموثوقية. ولمعالجة مشكلة عدم توازن فئات المشاعر، استُخدمت تقنيات Class Weights، وSMOTE، و Random Oversampling. وتم تقييم أداء مجموعة من نماذج تصنيف المشاعر شملت خوارزميات تعلم آلة تقليدية، ونموذج LSTM، إلى جانب نماذج لغوية عربية متقدمة مثل AraBERT و MARBERT. أما اكتشاف الموضوعات، فقد نُفذ باستخدام استراتيجية مزدوجة تجمع بين نموذج LDA الاحتمالي والتوليد القائم على GPT، بما يتيح فهمًا أعمق للمضامين الدلالية في مراجعات المستخدمين.

أظهرت النتائج تفوق النماذج اللغوية العربية المعتمدة على المحولات، حيث حقق نموذج MARBERT أعلى أداء بفضل قدرته على فهم السياق واللهجات العربية. كما بينت نتائج تحليل الموضوعات أن التعليقات الإيجابية تركز بشكل أساسي على سهولة الاستخدام وسلاسة تنفيذ المعاملات، في حين ترتبط التعليقات السلبية بمشكلات الاعتمادية التقنية، مثل تعثر تسجيل الدخول، وعدم استقرار التحديثات، وبطء الأداء، وتعطل التطبيق. وتوصي الدراسة بإعطاء أولوية لتحسين موثوقية التطبيقات المصرفية، وتعزيز اختبارات ما قبل الإطلاق، واعتماد آليات تحليل دوري لمراجعات المستخدمين باستخدام تقنيات معالجة اللغة الطبيعية لدعم تحسين جودة الخدمات.