Hierarchical Few-Shot Learning for Bilingual User Experience Classification: A Case Study on Saudi Arabia's Sehhaty Health App

Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
XX(X):1–17
c The Author(s) 2025
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Hissah.S.Al-Atwi [1], ;Areej.M.Al-Wabisi [1]; Salmah.M.Al-Qarni[1]; Gadaa.S.Al-Atwi [1] ;Wejdan.B.Al-Marwani[1] ;Omar I. Asiri[2], Fathi.A. Mubaraki[2]

[1] **Postgraduate Student (College of Computers and Information Technology) ; University of Tabuk: Tabuk, SA**
[2] **Assistant Professor (Faculty of Computer and Information Technology); University of Tabuk: Tabuk, SA**

ABSTRACT

This study examines user experiences in digital health applications through a hierarchical text-classification framework developed based on a healthcare- specific taxonomy. User reviews of the Sehhaty mobile application were collected from Google Play between 2019 and 2025 and analyzed bilingually in Arabic and English. After data cleaning and preprocessing, a corpus of more than 120,000 reviews was represented using the term frequency–inverse document frequency TF-IDF method.The hierarchical analytical framework organized user feedback into main themes including usability, performance, and support and corresponding subthemes reflecting detailed aspects of user interaction. Two supervised learning models were employed: a margin-based linear classifier for subtheme categorization and a probabilistic regression model for sentiment classification, both configured under few-shot learning conditions using limited labeled data. The thematic classifier achieved 99% accuracy, while the sentiment model obtained an F1- score of approximately 88%, demonstrating robust multilingual performance.Beyond textual features, emoji representations were visualized and analyzed as supplementary affective indicators within the perception layer of the analytical framework. Their inclusion enriched the interpretation of emotional patterns by revealing implicit user sentiments often overlooked in textual reviews. Temporal analysis revealed fluctuations in user satisfaction, with increased negative sentiment during the COVID-19 period due to greater reliance on remote healthcare. Overall, the findings highlight the significance of linguistically grounded computational modeling enriched with hierarchical and multimodal features for systematic monitoring and evidence-based improvement of national digital health infrastructures.

KEYWORDS: Mobile health apps, User Feedback Analysis, Cross-Lingual Sentiment Analysis, Few-Shot Learning Techniques, User Experience Enhancement

1. Introduction

Mobile health applications (mHealth) have come to represent indispensable tools to improve patients' access, offer efficiency and increase patients' engagement in healthcare services throughout the continuum of care according (Syukron *et al.*, 2023). Their importance grew considerably during the COVID-19 pandemic, where medical organisations came to rely heavily on digital platforms to provide continuity of medical supply and services through the intermediation of functionality such as appointment scheduling, access to electronic health records, preventive alerts, and telemedicine consultations (Alanzi, 2021). In Saudi Arabia, the Sehhaty mobile application has become the principal national portal to government-conceived health services and a principal enabler of the digital transformation vision in the health arena (Alanzi, 2021).

User experience (UX) plays an important part in the success and sustainability of digital health applications' user experience, which influences users' satisfaction, acceptance and continuous engagement and efficacy (Cassieri et al., 2024). An attractive, useful or otherwise user-friendly interface will enhance the user experience more effective than will the mere enlargement of information supplied from the application as stated by (Saoane Thach, 2019). Conversely, and negatively, technical difficulties such as slowness of application, logging on failings or system error can deter users from access to digital services, which will reduce trust in the provisions of those services by the system. Hence, to look at real-world user experiences through online commentary or review has established itself as one of the more effective approaches to identify usability difficulties and hurdles or areas of difficulty and a need for improvement, as supported by previous analyses of app-review datasets (Al Kilani, Tailakh & Hanani, 2019; Khan, Majumdar & Mondal, 2025).

Despite the increasing adoption of digital health platforms in the Kingdom of Saudi Arabia, there still exists a lack of comprehensive research into a full-blown analysis of large scale real world user experiences focusing on the national health applications (Cassieri *et al.*, 2024). This paper addresses this by analysing the user experiences through feedback and detailing the usability difficulties and trends of satisfaction detected in the reviews of the Sehhaty derived from a large menu of 'users' interaction and commentary on it.

To the best of our knowledge, no previous research has conducted a bilingual large-scale analysis of more than 120,000 Arabic and English commentary or review of the Sehhaty application on Google Play from 2019 to 2025. Existing studies have generally examined smaller datasets or focused on a single language (Alosaimi et al., 2024). Results derived from this study should provide greater insights into UX patterns which will identify recurring design and technical problems, which may result in improved user engagement and hence improve digitally offered health services, quality laws in alignment with the Kingdom of Saudi Arabia's transformed health objectives is essential in this research (Atchadé and Tchanati P., 2022).

Study Objectives

The goal of this work is to create a few-shot multilingual sentiment analysis model for Arabic and English user input in order to improve Saudi mHealth experiences. Motivated by the scarcity of large-scale bilingual investigations on Saudi national mHealth platforms and the absence of computational UX analysis tailored to Arabic user feedback, this study develops a bilingual hierarchical analytical framework to classify and analyse Arabic and English user reviews of the Sehhaty application—capturing themes, subthemes, sentiment, and experiential trends using few-shot machine-learning techniques. This approach addresses a research vacuum in healthcare sentiment analysis by providing multilingual few-shot methods with limited Arabic data, which is especially critical given Sehhaty's central role in national digital health transformation and the need to understand real user challenges, satisfaction trends, and priorities to guide evidence-based service improvement.

2. Related Work

Recent academic literature on the topic of digitalization in health services sectors, or in reference to Artificial Intelligence (AI) in Telehealth, has been a recognized innovation to transform and modernize the health system . According to(Amjad, et al, 2023), telehealth innovation has increasingly relied on artificial intelligence to improve healthcare services. In this context, m-health apps are a primary foundational consideration in this era in which high User Experience (UX) quality including consideration for adoption and effective clinical outcomes are paramount. The rapid reviews of the research reveal the importance of establishing crucial links between User Experience and Cybersecurity (Security) as success and sustainability criteria for digital health applications (Cassieri *et al.*, 2024).

It is suggested that App Review Mining is a powerful method for both qualitative and quantitative evaluation of UX after deployment, as user reviews contain valuable Software Requirements Engineering information, including improvement requirements related to usability or performance issues, as demonstrated in the study by (Tavakoli et al., 2018). Automatic categorization plays a crucial role in identifying evolving user needs and confirming recurring issues within mobile health application usage. It allows developers to interpret large volumes of user feedback in a structured manner, thereby transforming subjective user expressions into actionable insights that support continuous improvement and user-centered design. This importance becomes even more evident in healthcare contexts, where usability and reliability directly influence user trust and long-term engagement. As highlighted by(Al Kilani, Tailakh and Hanani, 2019), such categorization enhances the effectiveness of evaluation processes and contributes to more responsive application development. The relevant methodology and tools as an example MARK are available, providing sustainable health app UX by being able to monitor adverse keywords exposing a degradation in UX quality or operational performance overtime (Vu *et al.*, 2015). Systematic reviews have also referred to other research in order to notice gaps areas of UX in mobile applications has been reported on a different dimension and considered different methods of approaches to the literature every area of user experience will also not offer a suggested body of work from an original dimension of user experience development (Syukron *et al.*, 2023).

From a methodological perspective, researchers are interested in building Machine Learning (ML) models to effectively assess sentiment analysis from texts to determine user satisfaction. Supervised Algorithms are a fundamental consideration under the umbrella of the overall sentiment assessment process. In addition to sentiment analyses reliant on traditional statistical algorithms such as Naive Bayes in the case of work on sentiment analyses of patient feedback relating to the National Health Insurance Mobile Application in a health context based on the findings in their research (Kustanto, Nurma Yulita and Sarathan, 2021). Arabic sentiment modeling exercises are a high priority. With respect to the examples of Arabic text modeling a more advanced method of sentiment classification including measurement of user satisfaction with governmental health services apps relied on ML and Feature Engineering for Arabic texts. The relevant approach used the Support Vector Machine (SVM) algorithm with both Term Frequency-Inverse Document Frequency (TF-IDF) values for extraction of features from Arabic texts articulated proof of enhanced classification performance (Hadwan *et al.*, 2022).

Finally, in an additional reflective value there is a hybrid Arabic input model termed ArabBert-LSTM which incorporates gTransformer models with Long Short-Term Memory (LSTM) networks in an effort to advance efficacy in sentiment analysis of Arabic text (Alosaimi *et al.*, 2024). Furthermore, it is necessary to include Emojis in sentiment analysis with significantly advanced Transformer-based models, which reveal some subtle intents that were not possible with text only - However, it argues that sentiment analysis is essentially a user experience (Khan, Majumdar and Mondal, 2025). This user experience perspective is also directly reflected in adoption research, especially in the Saudi context, as empirical research indicates health information seeking and social influence were two of the main drivers to adopting mobile health apps in Saudi Arabia (Aljohani, 2025). Sentiment analysis has also been indicated of user comments in regards to COVID-19 applications in the Kingdom of Saudi Arabia, like `Sehhaty' and `Mawid', which comments consisted of hundreds of reviews to assess user's perception of the apps usability; which confirmed the focus of user experience QUIS (Alghareeb et al , 2023).

3. METHODOLOGY

This section describes the procedures used in this study, from collecting and cleaning the data, to preparing it for analysis, and finally analyzing it through successive analytical stages. Figure 1 presents sample screenshots from the Sehhaty application and its Google Play Store page, illustrating the context and source of the user-generated reviews analyzed in this study. This section seeks to be transparent in describing a process that is structured and repeatable and also accounts for potential bias while ensuring that the integrity of the data analyzed is fully retained throughout all stages.

Figure 2 outlines the overall methodological workflow used in this study, including data collection, cleaning, preprocessing, model training, evaluation, comparative analysis, and visualization.

The current study examines data on user reviews of the Sehhaty mobile application stored in the Google Play Store from 2019 to 2025. The dataset consists of 128,406 user reviews from the Google Play Store. After cleaning and preprocessing the data, the total reviews used were 120,561, of which 86,120 were in Arabic and 34,441 in English. The dataset can be found in the project's public GitHub repository.

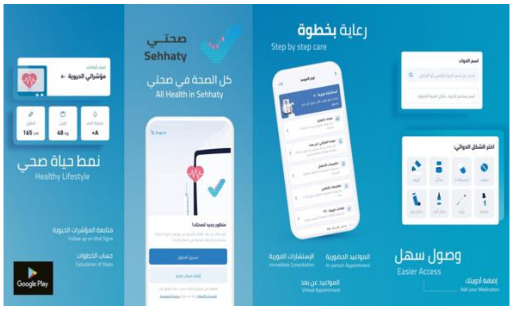

Figure 1. Sample screenshots from the Sehhaty app

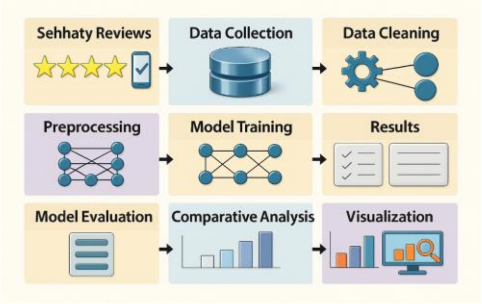

Figure 2. Workflow of the methodology undertaken in this study.

3.1 Data Cleansing

In this research study, data were extracted on July 25, 2025. A total of 128,406 unique user ratings for the Sehhaty mobile application were collected from the Google Play Store, over the period from 2019 to 2025. The data were subjected to standard cleaning procedures as a first step to enhance content quality before analysis. This also comprised the cleaning of ratings that were outside the 1–5 rating scale, removal of URLs, special symbols, repetitive characters, and diacritics. Transliterated Arabic language instances, written in Latin script, were also found and removed to maintain stylistic consistency across the dataset. At the same time, emojis remained since they are part of expressing sentiment; they are semantic indicators that can be integrated into sentiment analysis for better capture of emotional context and supporting proper classification of the text (Khan, Majumdar, & Mondal, 2025). To address redundancy, reviewer names were cross-checked against their corresponding comments. Names automatically generated by the system—such as "A Google user / مستخدم Google"—were disregarded during this step, as these identifiers are generic and cannot be used to infer duplication. Entries that appeared blank were revisited manually. During this step, any record showing content in languages other than Arabic or English was removed. Following the cleaning and refinement procedures, a total of 120,561 reviews remained suitable for analysis. Of these, 86,120 were written in Arabic and 34,441 in English. Overall, this finalized set accounted for 93.9% of all initially gathered records. At this point, entries that were duplicated or contained minimal or irrelevant content were excluded. This refinement contributed to a dataset that was clearer and more suitable for analysis. Figure 3 highlights the change in review numbers before and after the cleaning procedures, whereas Figure 4 presents the final language distribution of the dataset.

3.2 Ethical Considerations

The data used in this study consisted exclusively of publicly accessible user reviews obtained from the Google Play platform. No personally identifiable information was collected beyond the usernames automatically displayed by the platform. Phone numbers that appeared incidentally within some reviews were removed to ensure privacy. As the study relied solely on publicly available secondary data and did not involve any direct interaction with human participants, formal ethics approval was not required. All data collection and handling complied with Google Play's terms of service. This study used publicly available data and did not involve human subject.

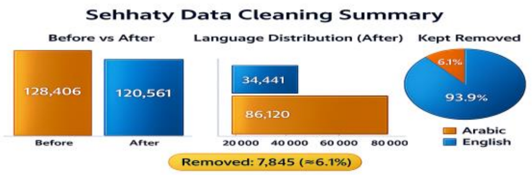

Figure 3. Comparison of the number of Arabic and English reviews before and after cleaning the data.

Dataset Language Breakdown
Arabic ≈ 71%
English ≈ 29%
Arabic
English

Figure 4. Language composition of the final dataset after cleaning. The 3D plot shows the portion of Arabic (~71%) and English (~29%) reviews after the data cleaning has finished

Figure 5. Arabic and English Word Clouds Before and After Text Cleaning

The initial exploration of the textual data through word clouds

Figure 5 visually presents the most frequent terms that appeared in both Arabic and English users' reviews before and after text cleaning. The clouds before cleaning have a lot of noise, namely repeated symbols, different spellings, and fragmented word forms. On the other hand, once the cleaning procedures are applied, the dominant terms then become more coherent and linguistically meaningful, enabling the representation of user sentiment and recurring concepts more clearly. Complementing these visual patterns, the most frequent unigrams, bigrams, and trigrams were subject to quantitative examination and were summarized, respectively, in Table 1, Table 2, and Table 3. These tables confirm the same patterns observed in the word clouds by identifying the most recurring single words, two-word expressions, and three-word expressions used by users to describe their experiences with the application.

Table1. **Most Frequent Unigrams in Arabic and English User Reviews**

Unigrams— Top 5			
Arabic	**Count**	**English**	**Count**
ممتاز	30,062	good	14,595
جدًا	12,717	app	6,126
التطبيق	5,711	very	4,760
جيد	5,408	nice	3,251
رائع	4,277	not	3,127

Table2. **Most Frequent Bigrams in Arabic and English User Reviews**

Bigrams — Top 5			
Arabic	**Count**	**English**	**Count**
ممتاز جدًا	5,290	very good	2,439
جيد جدًا	1,702	not working	1,082
لا يعمل	1,361	nice app	602
تطبيق ممتاز	1,200	very nice	525
جميل جدًا	1,153	app not	449

Table 3. **Most Frequent Trigrams in Arabic and English User Reviews**

Trigrams — Top 5			
Arabic	**Count**	**English**	**Count**
التطبيق لا يعمل	804	very good app	379
اكثر من رائع	320	app not working	269
تطبيق ممتاز جدا	265	not working after	133
شكرا وزارة الصحة	167	very nice app	101
تطبيق رائع جدا	160	not working properly	94

3.3 Data Preprocessing

Following the data cleaning process, additional preprocessing steps were applied to standardize the representation of the text such that it may be suitable for feature extraction and modeling. The primary processing steps included:

- Stop words removal: Custom stop lists were used for each language to retain evolutionarily common greetings and filler expressions.
- Spelling correcting: Short reviews were kept and standardized for frequent spelling corrections (common_fix) for both Arabic and English (e.g., → رايع رائع; dont → don’t).
- Handling missing values: The rest of the empty cells were standardized to NA.

Text normalization: All the preprocessing steps would normalize textual representation in both languages so the text could be represented uniformly for tokenizing and encoding later in the analysis.

Summary Tables 4 and 5 provide illustrative examples of these preprocessing steps, demonstrating how stopword removal and spelling correction were applied across both languages.

Table 4. **Examples of Stopword Removal**

✗ indicate that the word was removed during preprocessing (stopword deletion).

Examples of Stopword Removal

Language	Original	After	Status	Action
Arabic	**الخ**	**(removed)**	✗	**Delete**
English	**a**	**(removed)**	✗	**Delete**

- ### 3.4 Annotation Process

Before model development, a hierarchical taxonomy framework was created in order to guide the text classification process. The framework was developed based on a review of the relevant literature and an exploratory examination of a sample of user reviews to identify the most prominent themes and subthemes. Five main themes were deduced: User Experience & Sentiment, Technical Performance, Content & Services, Security & Support, and Suggestions & UI Design—all mapped to a more granular set of subthemes. Following the development of the taxonomy outlined above, 1,500 reviews (1,000 Arabic and 500 English) were manually annotated according to the predefined framework by following a few-shot learning strategy whereby, for each review, a main theme, subtheme, and sentiment category were assigned. Iterations took place in refining the labeling guidelines to support consistent interpretation. No formal inter-annotator agreement statistics were calculated, since the annotation process relies on a unified taxonomy and jointly reviewed guidelines that have been refined to ensure consistent interpretation and minimize variation among those carrying out the classification. The annotated subset provided the ground truth for subsequent model training and evaluation.

Table 5. **Examples of Spelling Correction (common_fix dictionary).**

✓ indicates that the word was corrected using the common_fix dictionary.

Examples of Spelling Correction (common_fix dictionary)

Language	Original	After	Status	Action
Arabic	**جميله**	**جميلة**	✓	**Correct**
English	**bcz**	**because**	✓	**Correct**

3.5 **Feature Extraction and Model Training**

Features in this study were generated by applying the TF-IDF technique on text data, a well-established method for converting textual input into numerical vectors that express how informative each term is within the wider corpus. This representation is well suited for short, user-generated content and serves as an effective way to highlight meaningful vocabulary.

To capture both individual token relevance and short contextual patterns, the TF-IDF vectorizerwas configured to include unigrams and bigrams (1–2 n-grams) for both Arabic and English texts. Two classical machine-learning models were subsequently trained on the TF-IDF features: a Linear Support Vector Classifier (Linear SVC) for subtheme classification and a Logistic Regression model for sentiment prediction. Training followed a few-shot configuration in which 1,000 Arabic and 500 English records were manually annotated for model development, while 24,000 records were reserved for testing (20% of the full dataset). The labelled subset size (1,500 records) was intentionally kept small to align with the principles of few-shot learning, which support effective generalisation with minimal annotated data (Song et al., 2022).

Classical machine-learning models were prioritised over deep-learning alternatives because they are more suitable when annotated data are limited, entail lower computational requirements, and provide greater interpretability—an important factor in health-related applications. Prior comparative work has also demonstrated that, under low-shot learning conditions, classical approaches such as SVM can perform on par with deep architectures (Alamri et al., 2024; Usherwood & Smit, 2019).

A fixed random seed (13) was used to support reproducibility. Theme labels were later inferred from their corresponding subthemes using a canonical mapping based on the study's taxonomy. When a review lacked meaningful free-text content, a fallback rule was applied to assign sentiment based on its numeric rating. This rule was applied to approximately 6.1% of all records, corresponding to entries without textual content but containing valid numeric ratings. Model implementation and optimisation were carried out using scikit-learn.

Summary Tables 6, 7, and 8 provide detailed descriptions of the training sample distribution, configuration parameters, and model–feature settings used in this study.

Table 6. **Training Sample Distribution by Language.**

Training Sizes	
Number of Labeled Samples	**Language**
Arabic	**1,000**
English	**500**

Table 7. **Training Sample Distribution**

General Configuration	
Parameter	**Value**
Total Samples for Texts	**24,000**
Random Seed	**13**

Table8. **Models and Feature Configuration**

Models and Features	
Category	**Model / Method**
Subtheme	**Linear SVC (TF-IDF 1–2)**
Sentiment	**Logistic Regression (class_ weight = "balanced")**
Theme	**Derived from Subtheme using the Taxonomy**

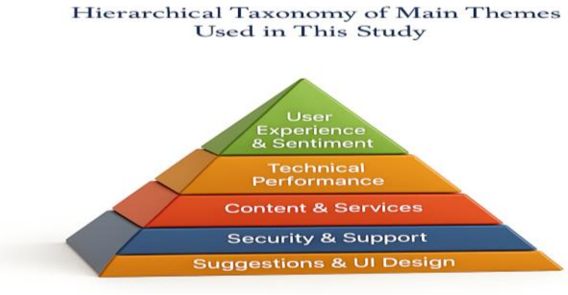
Hierarchical Taxonomy of Main Themes Used in This Study
User Experience & Sentiment
Technical Performance
Content & Services
Security & Support
Suggestions & UI Design

Figure 6. Hierarchical taxonomy used to construct themes and subthemes in this study

Table 9. **Final taxonomy used to develop thematic and sub-thematic classifications for user reviews**

Main Theme → Subthemes	
Theme	**Subthemes**
User Experience & Sentiment	**Ease of Use, Navigation, UI Clarity, Onboarding, Overall Satisfaction, Accessibility, Help & Guidance, General, General_UX**
Technical Performance	**App Speed, Loading Time, Crashes/Freezes, Errors/Bugs, Connectivity/Network, Stability, General, General_Technical**
Content & Services	**Appointment Booking, Results Delivery, Reports/Documents, Prescriptions, Records/Vaccination, Teleconsultation, General, General_Content**
Security & Support	**Login/OTP, Password Reset, Account Verification, Privacy/Permissions, Support Responsiveness, Account Access Issues, General, General_Security**
Suggestions & UI Design	**Feature Request – Dark Mode, Notifications & Reminders, Layout Improvements, Customization, Language Options, Accessibility Enhancements, General, General_Suggestions**

Figure 6 visualises the hierarchical taxonomy used to organise the main themes and subthemes referenced during model training, while Table 9 presents the final structure of thematic and sub-thematic categories derived from this taxonomy for the classification tasks. In line with this structure, general subtheme labels, such as General_UX, General_Technical, General_Content, General_Security, and General_Suggestions, were retained to categorize non-specific user reviews that aligned with the main theme but lacked sufficient detail to be assigned to a more precise subtheme.

3.6 Evaluation

In the model assessment phase, a stratified K-Fold cross-validation procedure was adopted to examine how consistently and accurately the models performed. By keeping the class proportions similar across folds, this method helped reduce evaluation bias. Roughly 24,000 records—about 20% of the dataset—were used and split into five folds to support balanced testing. Several macro-based performance indicators were applied. Accuracy represents how many samples were correctly predicted. Macro-precision reflects the proportion of predicted positives that were correct, while macro-recall indicates how many relevant samples were detected. The macro F1-score combines both measures to provide an overall performance indicator. Collectively, these metrics provided a structured understanding of the models' behaviour in multilingual text classification tasks. They also supported balanced learning, reducing the likelihood of both overfitting and underfitting, and enabling robust, reliable predictive performance.

4. Results

This section discusses the experimental results of the study, emphasizing the classification models' performance while noting the evaluation metrics used. The results indicate the findings from applying the method proposed through the methodology on the Arabic and English datasets of the Sehhaty mobile application. A thorough discussion of the key performance metrics of accuracy, precision, recall, and F1-score will also be included, notwithstanding a discussion of possible bias influences and class imbalance, to ensure we have considered an appropriate and unbiased interpretation of the results.

4.1 Overview

This section introduces the performance results of the few-shot classification models on Arabic and English user reviews of the Sehhaty mobile application. The results show the classification accuracy across Theme, Subtheme, and Sentiment using stratified K-Fold cross-validation (k=5).

4.2 Model Evaluation

To ensure fair measurement of model

performance, a 5-fold stratified cross-validation is used to balance the class distribution (approximately 80% belonged to positive sentiment). Accordingly, macro-averaged metrics (Precision, Recall, F1 score) were used for evaluation because these metrics do not account for the class frequency.

Table 10. **Performance metrics of the classification models using stratified K-Fold (k=5)**

Metric	Performance of few-shot models on Sehhaty dataset (k=5)		
	Theme	Subtheme	Sentiment
Accuracy	99.70%	99.91%	**99.93%**
Precision	93.99%	95.30%	**87.26%**
Recall	97.81%	97.16%	**89.97%**
F1-Score	**95.61%**	**96.03%**	**88.20%**

These results show that all three levels of classification perform very well and consistently, with very high accuracy values across the models. In the few-shot setting, precision and accuracy for Theme, Subtheme, and Sentiment classification resulted in values higher than 99%, thus showing stability and reliability for the chosen modelling approach even with minimal labeled data. The F1-scores balancing precision and recall were situated between 88% and 96%, suggesting a good generalization of models across Arabic and English reviews. Among the three classification layers, the highest F1-score was achieved by the Subtheme classifier, 96.03%, which reflects that the TF-IDF representation combined with the Linear SVC classifier was particularly effective for grasping the fine-grained contextual patterns in users' comments. The Sentiment classifier recorded somewhat lower precision and recall, although its performance was still robust. This is perhaps due to mixed emotional expressions and bilingual phrasing that exist in user reviews. The class imbalance probably accounted for the strong accuracy in sentiment prediction to some extent, given that some 80% of reviews had positive sentiment, making it much easier for the model to accurately predict the majority class. Table 10 summarizes these performance results by providing a summary of the accuracy, precision, recall, and F1-scores for the Theme, Subtheme, and Sentiment classification tasks. These results together suggest that the modelling framework applied performed reliably across all metrics of evaluation, with the classifier for Subtheme providing the best overall balance between precision and recall.

4.3 Comparative Analysis

In this section, we will be comparing the performance of our proposed few-shot classification model against the performance of earlier studies that have conducted sentiment analysis on Arabic datasets. Table II summarizes the models, dataset types, and performance metrics reported in the relevant literature.

Table 11. **Comparative results of previous studies and the proposed few-shot model**

Study	Comparative performance of classification models			
	Accuracy	F1-Score	Recall	Precision
Hadwan et al.	95.1%	99.91%	95.0%	**93.5%**
Alsemaree et al.	96.8%	95.9%	96.8%	**95.9%**
Alosaimi et al.	97.7%	96.6%	97.7%	**97.2%**
Present Study	**99.7%**	**95.6%**	**97.8%**	**93.9%**

The comparative results presented in Table 11 clearly demonstrate that the proposed model performs substantially better than, or at least on par with, previous studies that conducted sentiment or thematic classification on Arabic datasets. With an overall accuracy of 99.7% and a strong recall of 97.8%, the model has shown an exceptional ability to correctly identify class labels across a diverse range of user feedback while maintaining stability across all evaluation metrics. These results emphasize the model’s robust generalization capability, particularly in healthcare-related applications where multilingual and unstructured user input is common.

Furthermore, the consistency of performance across theme, subtheme, and sentiment classification indicates that the adopted TF-IDF and Linear SVC approach remains highly effective even with limited labeled data, reinforcing the findings summarized in Table 11.

Table12. **Summary of model architectures and dataset characteristics in comparable studies**

Study	Study Details			
	Model	Dataset Type	Language	Year
Hadwan et al.	TF-IDF + SVM	App Reviews	Arabic	**2022**
Alsemaree et al.	TF-IDF + Ensemble ML	Twitter Posts	Arabic	**2024**
Alosaimi et al.	Arab BERT + LSTM	Text Corpus	Arabic	**2024**
Present Study	**TF-IDF + Linear SVC / LogReg**	**App Reviews (Sehhaty)**	**Mixed (AR + EN)**	**2025**

Unlike most prior studies, which applied binary sentiment analysis, this study employed a three-level classification structure comprising Theme, Subtheme, and Sentiment. This nested design enabled a deeper and richer interpretation of user feedback by capturing both topical categories and emotional tone, consequently providing more contextual insight than was the case with sentiment analysis alone. Table 12 compares the model architecture and dataset characteristics in similar studies and highlights how the current one differs by incorporating reviews in Arabic and English. Figure 7 further shows that the proposed framework of TF-IDF + Linear SVC/LogReg outperforms and sometimes outcompetes more complex neural frameworks employed in previous studies. The multilingual nature also meant that the design could handle diverse user-generated content, which is crucial in health applications and serves a heterogeneous population. Whereas previous studies generally adopted neural architectures such as BERT-based models, the current results suggest that equally strong accuracy, F1-scores, recall, and precision can be obtained at a fraction of the computational cost using a classical lightweight approach. That characteristic itself will be advantageous for practical deployments in healthcare settings where computational efficiency and interpretability are key demands. Overall, these findings highlight the efficiency, robustness, and practicality of the few-shot model proposed, showing that a classical TF-IDF–based linear architecture can rival its neural counterparts in accuracy and generalization, while retaining interpretability—a potential advantage in real-world healthcare applications.

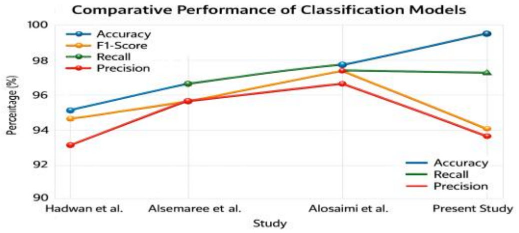

Figure 7. Comparative Performance of Classification Models

4.4 Visualization of Analytical Findings

To complement the quantitative evaluation, this section offers a series of visual analyses that outline the key findings revealed by the dataset. The graphics portray complicated analytic findings in a clear and succinct manner, showing the general sentiment trends, distributions of ratings and criticisms, technical issues, and user engagement actions. Relationships between the content of the review, the language of the review, and the user experience are revealed through these visual summaries, permitting further interpretation of the behavioral and thematic aspects uncovered in the study. Each figure presents a specific analytical focus which reinforces the major findings of the study.

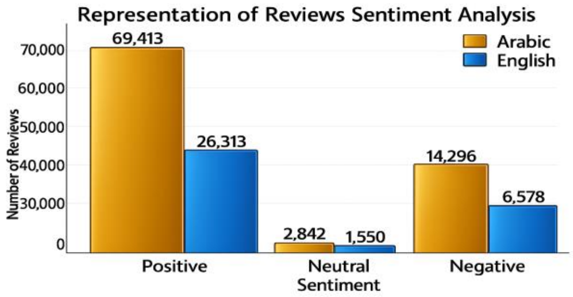

Figure 8. Comparative Sentiment Distribution of Arabic and English User Reviews in the Sehhaty Dataset

The overall distribution of sentiment for all user reviews for the Sehhaty application shows a greater occurrence of positive feedback (80.6% Arabic; 76.4% English), which indicates a general acceptance of the services and usability of the application. There are very few neutral comments, and there also exists a minority of negative comments (16.1% Arabic; 19.1% English) that can be attributed to a lack of access, e.g., login difficulties and slow responses. Overall, this shows the consistent appreciation of the platform on the part of the user for the accessibility of the platform and repeated cries for easier appointment scheduling, plus quicker response times. Figure 8 visually supports this distribution by clearly illustrating the predominance of positive sentiment across both languages.

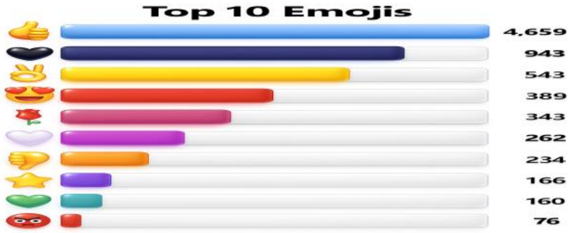

Figure 9. Distribution of the Top 10 Most Used Emojis in Sehhaty App Reviews

Figure 9 illustrates the distribution of the ten most frequently used emojis in Sehhaty app reviews, providing a clear visual summary of which symbols appeared most commonly across user comments. This figure shows the ten most used emojis in comments that were provided by users, indicating which symbols appeared most across the reviews. From 2019 to 2025, respondents frequently used emojis to give an opinion about the Sehhaty app. These visual cues reflected a number of different emotions—both positive and negative—and indicated a general sense of the opinions that users formed of the service and their level of engagement with it. They were hardly random decoration; sometimes, emojis expressed feelings where text did not. Attending to them allows one to discover reactions that users do not articulately mention, while overlooking them would only result in losing critical insights into their experiences altogether (Khan, Majumdar, & Mondal, 2025).

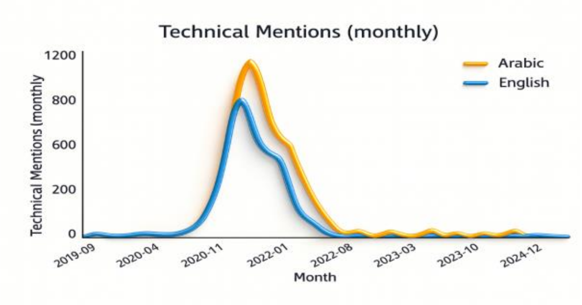

Figure 10. Comparative temporal trend of technical issues in Arabic and English user reviews

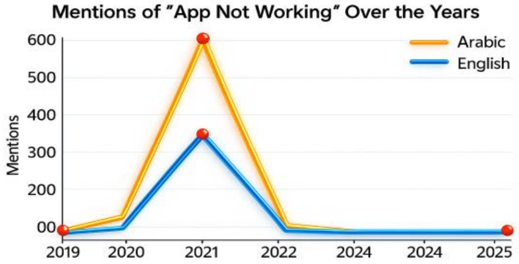

Figure 11. Comparative trend of "App not working" mentions in Arabic and English user reviews.

The annual and monthly patterns of technical issues identified in both Arabic and English reviews of the Sehhaty application are illustrated in Figures 10-11. The most common technical problems identified referred to slow operation, crashes, or network problems and represented between 8%-11% of all user comments. A sharp increase in complaints occurred in 2021 which coincided with the COVID-19 pandemic period and in which user demand for health-related digital services increased. The following years exhibited a definite decrease in issues with this phenomenon indicating substantial improvement in the stability of the system and responsiveness of the app. Specifically, the phrase "App not working" appeared in the Arabic reviews 806 times and in the English reviews 307 times. Both of these phenomena also exhibited an increase in usage during 2021 and a gradual decrease thereafter. These data indicate the dynamic relationship which exists between app performance and external factors such as user load and public health events and indicate favorable results of technical optimization.

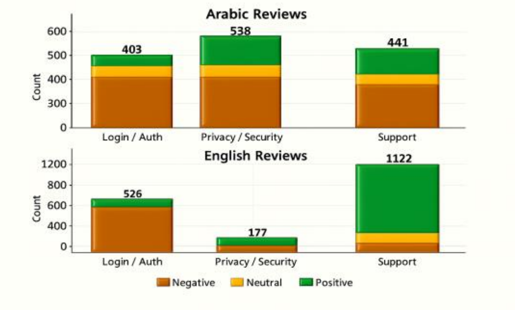

Figure 12. Sentiment distribution across Security and Support categories in Arabic and English datasets.

Figure 12 presents the sentiment distribution across Security and Support categories in both Arabic and English datasets. User feedback regarding Security and Support from both Arabic and English datasets has indicated that Arabic reviews mostly concentrated on Privacy/Security and Login/Auth aspects, being dominated by negative sentiments, reflecting mainly in log-in failures and privacy concerns. The English reviews are focused more on Support requests with special mentions of queries on help and verification etc. and exhibited a more even weather of sentiment distribution.

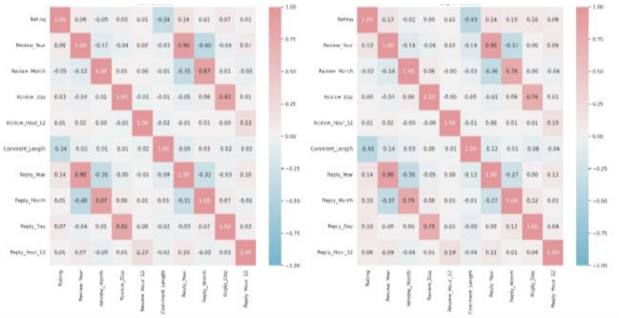

Figure 13. Comparative correlation patterns between Arabic and English user review Datasets

Figure 13 illustrates these correlation patterns across both datasets. The analysis revealed a very high positive correlation in respect of temporal variables in the Arabic and English data, e.g., Review year × Reply year (r = 0.90) and Review month × Reply month, indicating that replies are made in a regular temporal order. However, moderate negative correlations exist, e.g., in Rating × Comment length (r ≈ -0.4) in that shorter comments are likely to be more favourable and longer ones more unfavourable. Overall, it was shown that stable user–developer interaction takes place with small variations in time and different degrees of expressiveness.

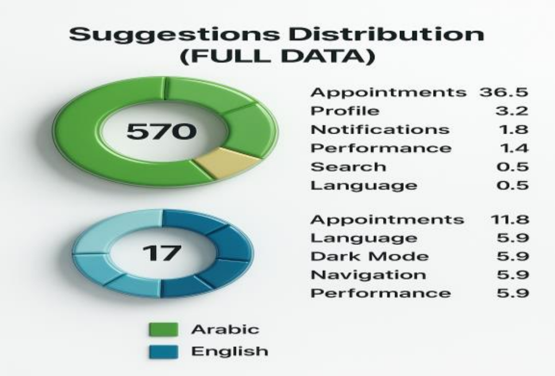

Figure 14. shows the distribution of user suggestions in Arabic and English datasets. The graph shows the distribution by category of user suggestions.

The Arabic reviews gave many more suggestions, which referred mainly to appointments while minor suggestions were made with regard to profiles, notifications, and performance. In English reviews, fewer suggestions were given. The most frequently given suggestions concerned appointments, language, and navigation. As illustrated in Figure 14, appointment-related suggestions constituted the largest portion of all recommendations, with approximately 36.5% of Arabic suggestions (out of 570) and 11.8% of English suggestions (out of 17) explicitly requesting improvements in appointment-booking processes.

5. Recommendations

Taking into consideration the results from the various analyses which show the effectiveness of the use of hierarchical taxonomy structure for uncovering main and sub themes, rather than giving a general performance of the sentiment classification, and due to the results gained from visual analyses which have shown patterns regarding the satisfaction and discontent of users, give way to the following recommendations for exploitation in the future.

1) Aspect Based sentiment analysis: in fact, the sentiment analysis should be carried out with regards the aspect rather than in terms of the whole review, targeting the sub theme of such as performance, usability, and appointments, this allows the placing of the positive and negative sentiments detected within each dimension of the user experience.

2) Improvement of Performance of Sentiment Model to improve the performance of the classification system of the sentiment analysis it is recommended to introduce means so that the problems regarding the imbalance of classes may be dealt with it is a better application of the strong Arabic transformer models such as arabert, and XLM-roBERTa.

3) Introduction of the Temporal aspect: This means should be input into the temporal aspect of the framework, so that research may be made in relation to the dynamic of the sentiment and themes in relation to the updates of the app or changes of service that the trends regarding the user's satisfaction may be followed.

4) Co-occurrence analysis of themes: It is suggested that the investigation of the types of themes which when mentioned in the user reviews frequently co-occur with each giving rise to possible correlations in the categories usability, performance and privacy, support etc.

5) Creation of a Dashboard of an Interactive Nature: It is desirable to create the results gained from the analyses of these in the nature of dashboard which is of an interactive nature and are in a state of continual updates

whereby the user has a close how user satisfaction is, and the problems identified by the sentiments of users are indicated in real time.

6) Extension of Hierarchical Taxonomy**:** All machine learning methodologies available should be in a state of continual update and enrichment of the taxonomy in a way that hidden patterns of user expression are gained which are revealed with the passage of time which will allow the system to continue and be continually and full of the correct contextual information.

6 . Conclusion

This study provided an extended evaluation and analytic method to describe the user experience of the Sehhaty application from a general position within the complete context of digital health care services. This was aided by a developed taxonomy driven hierarchical classification allied with a method of sentiment analysis. The specific approach undertaken in this study which used a method of analysis known as Few-Shot Learning, allowed for a user experience classification to be made in an operationally efficient method even though there were not enough labelled data available.

The importance of using systematic methods of classification in order to create meaningful categories from unstructured user feedback regarding digital health care services was demonstrated by this result. The important conclusion of this study was that the use of linguistic analysis methods in conjunction with hierarchical taxonomic classification and Few-Shot Learning leads to an additively greater efficacy in the analysis from user comment opinions regarding the Sehhaty application. The study can be seen as a logical extension to the present state of the art in digital health analytics since it proposes a scalable solution to be created that could provide the means to assess the user satisfaction of other health care applications. Overall, it can be seen that the study results show the necessity of a continual development of the Arabic language models and related hierarchical taxonomy system, and that of interactive methods of analysis to allow a satisfactory transition of the findings of the study into tangible results. In addition to this the combination of the hierarchical taxonomy and Few-Shot Learning provides a new method for applying this means to other fields of interest which rely on a small degree of text information transference to promote artificial intelligence towards the end of facilitating sustainable methods of innovation within the health care industry.

8 . Data & Code Availability

All code, labeled sample data, and project documentation used in this study are publicly available in the following GitHub repository:

GitHub Repository:
https://github.com/hissah055/sehhaty-reviews-ml-2025/tree/main

References:

1. Al Kilani, N., Tailakh, R. and Hanani, A. (2019) 'Automatic Classification of Apps Reviews for Requirement Engineering: Exploring the Customers Need from Healthcare Applications', in 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain: IEEE, pp. 541–548. Available at: https://doi.org/10.1109/SNAMS.2019.8931820.
2. Alamri, Alamri, A., Al-Ahmadi, A., Al-Ghamdi, M. and Al-Khalifah, T. (2024) 'Benchmarking Classical vs Deep Learning Methods for Arabic Short-Text Classification Under Limited Resources',

Heliyon, 10(4), e25612. Available at: https://doi.org/10.1016/j.heliyon.2024.e25612

3. Alanzi, T. (2021) ‘A Review of Mobile Applications Available in the App and Google Play Stores Used During the COVID-19 Outbreak’, Journal of Multidisciplinary Healthcare, Volume 14, pp. 45–57. Available at: https://doi.org/10.2147/JMDH.S285014.
4. Alghareeb, M., Albesher, A.S. and Asif, A. (2023) ‘Studying Users’ Perceptions of COVID-19 Mobile Applications in Saudi Arabia’, Sustainability, 15(2), p. 956. Available at: https://doi.org/10.3390/su15020956.
5. Aljohani, N. (2025) ‘Digital Health Transformation in Saudi Arabia: Examining the Impact of Health Information Seeking on M-Health Adoption during the COVID-19 Pandemic’, Engineering, Technology & Applied Science Research, 15(1), pp. 19933–19940. Available at: https://doi.org/10.48084/etasr.8747.
6. Alosaimi, W. et al. (2024) ‘ArabBert-LSTM: improving Arabic sentiment analysis based on transformer model and Long Short-Term Memory’, Frontiers in Artificial Intelligence, 7, p. 1408845. Available at: https://doi.org/10.3389/frai.2024.1408845.
7. Alsemaree, O. et al. (2024) ‘Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions’, Heliyon, 10(9), p. e27863. Available at: https://doi.org/10.1016/j.heliyon.2024.e27863.
8. Amjad, A., Kordel, P. and Fernandes, G. (2023) ‘A Review on Innovation in Healthcare Sector (Telehealth) through Artificial Intelligence’, Sustainability, 15(8), p. 6655. Available at: https://doi.org/10.3390/su15086655.
9. Atchadé, M.N. and Tchanati P., P. (2022) ‘On computational analysis of nonlinear regression models addressing heteroscedasticity and autocorrelation issues: An application to COVID-19 data’, Heliyon, 8(10), p. e11057. Available at: https://doi.org/10.1016/j.heliyon.2022.e11057.
10. Available at: https://arxiv.org/abs/1907.07543.
11. Cassieri, P. et al. (2024) ‘User Experience and Security in Digital Health Applications: Results from a Rapid Review’, in 2024 50th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). 2024 50th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Paris, France: IEEE, pp. 296–299. Available at: https://doi.org/10.1109/SEAA64295.2024.00053.
12. Hadwan, M. et al. (2022) ‘An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services’ Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique’, Applied Sciences, 12(11), p. 5547. Available at: https://doi.org/10.3390/app12115547.
13. Khan, A., Majumdar, D. and Mondal, B. (2025) ‘Sentiment analysis of emoji fused reviews using machine learning and Bert’, Scientific Reports, 15(1), p. 7538. Available at: https://doi.org/10.1038/s41598-025-92286-0.
14. Kustanto, N.S., Nurma Yulita, I. and Sarathan, I. (2021) ‘Sentiment Analysis of Indonesia’s National Health Insurance Mobile Application using Naïve Bayes Algorithm’, in 2021 International Conference on Artificial Intelligence and Big Data Analytics. 2021 International

Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA), Bandung, Indonesia: IEEE, pp. 38–42. Available at: https://doi.org/10.1109/ICAIBDA53487.2021.9689726.

15. Saoane Thach, K. (2019) ‘A Qualitative Analysis of User Reviews on Mental Health Apps: Who Used it? for What? and Why?’, 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1–4. Available at: https://doi.org/10.1109/RIVF.2019.8713726.
16. Syukron, F.W. et al. (2023) ‘Exploring User Experience in Mobile Applications: A Systematic Literature Review’, in 2023 11th International Conference on Cyber and IT Service Management (CITSM). 2023 11th International Conference on Cyber and IT Service Management (CITSM), Makassar, Indonesia: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/CITSM60085.2023.10455498.
17. Usherwood, P. and Smit, S. (2019) ‘Low-Shot Classification: A Comparison of Classical and Deep Transfer Machine Learning Approaches’, arXiv preprint, arXiv:1907.07543.
18. Vu, P.M. et al. (2015) ‘Tool Support for Analyzing Mobile App Reviews’, in 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE). 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), Lincoln, NE, USA: IEEE, pp. 789–794. Available at: https://doi.org/10.1109/ASE.2015.101.