Vol.3, No.1, Page 646- 663, 2025 https://doi.org/10.59628/jast.v3i1.1364 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

Utilizing Machine Learning based on LLM for Arabic Sentiment Analysis in Assessing User Satisfaction with Mobile Banking Apps: A Case Study of Yemeni Banks

Salah AL-Hagree [1,2] * and Ghaleb Al-Gaphari[2]

[1]**Department of Computer Science , Faculty of Computer and Information Technology, University of Sana'a, Sana'a, Yemen,**
[2]**Department Computer Science, Faculty of Sciences, University of Ibb, Ibb , Yemen**

*Corresponding author: s.alhagree@su.edu.ye

Abstract

The development of large language models (LLMs) that are optimized to obey human commands is a significant advancement in the field of artificial intelligence (AI). One such model is ChatGPT (Chat Generative Pre-trained Transformer) from OpenAI, which has shown itself to be an extremely powerful tool for a variety of tasks such as conversation production, code debugging, and answering questions. Despite the fact that these models are praised for their multilingualism, little research has been done on how well they can analyze sentiment, especially in Arabic. We intend to close this gap by thoroughly assessing ChatGPT's sentiment analysis skills, particularly with regard to Arabic text, in light of this constraint. When developing applications, recognizing the quality of the application and satisfying user needs are essential. Understanding user requirements is crucial to improving the quality of programs. The use of application review-based sentiment analysis (SA) is one efficient method for accomplishing this. The purpose of this study was to evaluate consumer perceptions of mobile banking apps so that they could be updated and maintained appropriately. Since mobile banking apps are now a necessary part of people's lives, it is essential to examine user reviews of these apps for SA purposes. User reviews of banking mobile apps on the Google Play Store provided the dataset used in this study. We suggest a new active labeling technique for ChatGPT and examine the effects of using the ChatGPT variants for Arabic sentiment analysis (ASA). Using the accuracy, recall, precision, and F-score measures, we assess the performance of four machine learning (ML) approaches: Naive Bayes (NB), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and Random Forest (FR). Additionally, we contrast six approaches to data labeling: manual human labeling, ChatGPT labeling by Assistant-Poe, ChatGPT labeling by Bing-Edge, ChatGPT labeling by Assistant-Poe with humans, ChatGPT labeling by Bing-Edge with humans, and ChatGPT labeling by Assistant-Poe with Bing-Edge. Using different Bing-Edge models for ASA, our experimental results demonstrate that the NB approach performed the best, with an accuracy of 91.22%, recall of 89.62%, precision of 88.90%, and F-score of 89.26%. Additionally, we contrast two approaches of data labeling for ASA: human labeling and labeling with Bard Google. Using Bard Google models for ASA, our experimental results demonstrate that the Naive Bayes technique outperformed the others, attaining an accuracy of 98.07%. Furthermore, when compared to alternative labeling techniques, our suggested active labeling strategy with ChatGPT produced greater accuracy. Our research indicates that our suggested active labeling method and the NB technique with multiple Bing-Edge models are useful strategies for ASA using ChatGPT. Our research provides important insights into efficient methods for this endeavor and advances the field of sentiment analysis in Arabic literature. Additionally, compared to other labeling techniques, our suggested active labeling method using Bard Google produced greater accuracy. According to the proposed study, our suggested active labeling strategy and the Naive Bayes technique with Bard Google models are both efficient methods for Arabic sentiment analysis utilizing Bard Google.

Article Info

Keywords:
Arabic Sentiment Analysis, Banking Services, Bard Google, ChatGPT, Machine learning

Article History:
Received: 2-November-2024,
Revised: 23-December-2024,
Accepted: 15-January-2025,
Available online: 28 February 2025.

1. INTRODUCTION

Banks provide a variety of banking services to their customers, and they actively engage with customer feedback and address their concerns. Evaluating the impact of these banking services on our daily lives is crucial. One modern approach to assess customer feedback is through analyzing user reviews of banking service applications. This can be achieved through sentiment analysis (SA). SA is capable of uncovering individual opinions and sentiments regarding a particular topic. For example, it can help determine marketing strategies, business priorities, and product enhancements [1]. Before planning to launch a product, manufacturers need to know what potential customers expect. This is done to ensure that those goods will be what people want [2]. Prospective customers typically examine user reviews before deciding whether to utilize a product. The product supplier can then determine whether the product's quality can meet user needs [3]. SA is a way to gauge how satisfied customers are with a product. Sentiment analysis has a wide range of uses, including online brand monitoring and reputation management. Keeping an eye on your own communications, investment information, crime prevention, social media monitoring, and ticket analysis for customer service Pay attention to the employee's voice, the customer's voice (VoC), Market research, competitive research, and product analysis Purchasing a service or a commodity, In anticipation of elections, Enhancing a product or service's quality, making decisions, identifying unwelcome viewpoints, and identifying bullying or incitement. Arabic sentiment analysis is commonly employed to determine individual perspectives on various subjects, such as events, products, or other entities [3]. In sentiment analysis tasks, text records are categorized into three classes: positive, negative, or neutral, representing different levels of text polarity [4]. There are two primary approaches to sentiment analysis: lexicon-based sentiment analysis (LBSA) and ML Arabic sentiment analysis (MLSA). LBSA uses a vocabulary dictionary to calculate the polarity of each text record, whereas MLSA relies on ML models to predict the polarity of text records. Although MLSA is more efficient, it requires human-annotated data for training on polarity before the prediction process [5]. Customer experience (CX) analysis indicates that while the majority of CX solutions available on the market provide sentiment analysis in many languages, machine translation is regrettably the most used method. A CX software platform, such as Adobe Experience Clarabridge, Manager, IBM Tealeaf ClickTale OpenText, Medallia, Satmetrix, Qualtrics, Zendesk and Zoho CRM Plus، is a technological solution that assists businesses in managing, measuring, and enhancing the overall CX with the organization. While machines may translate some languages with similar roots, including Spanish, Italian, Latin, and English„ languages with entirely different linguistic laws cannot be translated in the same way. This is because translations fail to capture the intricacy of subtleties in languages as rich as Arabic [6]. The Arabic language is widely utilized and well liked. Therefore, it is critical to use sentiment analysis techniques when learning Arabic. However, the morphology, structure, and diversity of the Arabic language make this difficult. Sentiment analysis in Arabic still requires further work [7]. A particular Arabic dataset was gathered for this study and carefully annotated for SA tasks. Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and K-nearest neighbor (KNN) algorithms were used to conduct sentiment analysis in Arabic. What are some of the difficulties in analyzing sentiment in Arabic banking? The mobile applications of Yemeni banks were targeted. Four ML assessment metrics were used to evaluate the model. The advantage of the NB algorithm is revealed by comparing the four models. An important development in artificial intelligence is the creation of large language models (LLMs) that are tuned to follow human instructions. OpenAI's ChatGPT is one such model that has demonstrated exceptional performance in a range of tasks including code debugging, question answering, and dialogue synthesis. Although the multilingualism of these models has received much attention, nothing is known about how well they can assess sentiment, especially in Arabic. By carefully evaluating ChatGPT's sentiment analysis abilities in the Arabic language, we hope to bridge this gap and overcome this limitation. The goal of this endeavor is to better understand the capabilities of the model and its use in Arabic sentiment analysis [8]. Arabic text data have significantly increased as a result of the rapid development of social media and Google Play platforms, as well as the rise in online communication. Sentiment analysis is a popular method for examining web information and determining user attitudes and views about a certain subject. However, because Arabic is a complex language, sentiment analysis of Arabic text has always proven to be difficult. The use of sophisticated natural language processing (NLP) models, such as ChatGPT, to increase the accuracy of Arabic sentiment analysis is becoming increasingly popular [9]. The purpose of this work is to examine how ChatGPT, a cutting-edge NLP model, affects the Arabic sentiment analysis (ASA). Using a dataset of evaluations of Arabic mobile applications, this study assessed the performance of ChatGPT and compared its outcomes with those of other ML methods [7]. The results of this study will shed light on how well ChatGPT works for the ASA and demonstrate how it may be used to increase the precision of sentiment analysis of Arabic text. In this paper, we present the development of a model for Arabic Sentiment Analysis (ASA) using data labeling (DL) techniques, leveraging both human input and ChatGPT and employing ML methodologies such as NB, K-NN, SVM, and RF. With the rapid growth

of digital content in Arabic on the Internet, there is increasing interest in the SA of Arabic texts. The challenge of DL and annotation is a critical phase that significantly affects the accuracy and effectiveness of the SA models. We discuss various DL techniques and the importance of the quality of labeled data in the training models. This paper outlines the steps taken to develop and validate the model using a relabeled dataset based on ChatGPT and humans. The goal of this study is to provide a reliable model for SA, which can enhance the understanding of Arabic digital content and support various applications such as public opinion monitoring and user experience improvement. The main contribution of this study is maximizing the efficiency of the ASA technique by using the following:

- Creation of novel dataset for user comments collected from banking mobile applications (apps) on Google Play Store to perform ASA.
- Development of ML based ChatGPT to perform ASA.
- To compare the performance of humans and ChatGPT with other conventional SA techniques on the same labeled dataset.
- To determine how various labeling strategies affect ChatGPT's and humans' performance for ASA
- To examine how various pre-processing methods affect the precision of ASA tagging using ChatGPT and people.
- To investigate how people and ChatGPT can work together to improve ASA tagging accuracy and solve the difficulties associated with Arabic text analysis.
- To emphasize the possibility for further research in this area and offer insights on the efficacy of labeling for ASA based on people and ChatGPT.
- To the best of our knowledge, this work is the first to compare ChatGPT and humans on NLP tasks in many languages, with an emphasis on speech recognition (ASA).

The remainder of this paper is organized as follows. Section 2 provides an overview of recent studies on sentiment analysis that utilize app feedback data from banking mobile apps on Google Play Store, with particular emphasis on the Arabic language. Section 3 presents the proposed models and methods, including a brief description of the dataset and the preprocessing techniques employed. The findings and discussions of the experiments are presented in Section 4. Finally, Section 5 presents the conclusions of the study.

2. LITERATURE REVIEW

To identify the research gap in sentiment analysis, we conducted a thorough review of important and relevant studies. The practice of identifying patterns in textual data, such as classifying and interpreting sentiment into neutral, positive, or negative remarks using methods such as ML, is known as sentiment analysis. Sentiment analysis has emerged as a crucial technique for comprehending user beliefs and actions because of the growing amount of user data available online, including social networks and other platforms. The development of sentiment analysis tools based on Bard Google has been the subject of numerous studies. Furthermore, sentiment analysis assessments are frequently drawn from a variety of channels, including Google Plays. We hope to advance the field of ASA and offer a more precise and efficient method of sentiment analysis in Arabic text by investigating these studies and platforms. The difficulties of sentiment analysis in Arabic and the paucity of research on ASA in comparison to English and other Latin languages were examined in an earlier study [6]. Using a dataset of user comments from evaluations of certain mobile applications on Google Play Store, the study suggested a novel method for assessing sentiment in Arabic script. The strategy entailed enhancing data pretreatment algorithms, such as the Levenshtein distance (LD) algorithm, and merging it with the K-NN algorithm. The influence of effectively using the K-NN and LD algorithms for the ASA on mobile application reviews was examined through tests. According to the findings, the K-NN with the LD algorithm produced the best evaluation metrics for accuracy, recall, precision, and F-score. This study showed how the suggested method could increase the precision and potency of sentiment analysis in Arabic text. A prior study [7] sought to ascertain consumer perceptions of mobile banking service applications and maintain and enhance them. The study analyzed user feedback gathered from banking mobile apps on Google Play Store using application review-based sentiment analysis (SA). Three classes, positive, negative, and neutral, were manually assigned to the dataset. Arabic sentiment analysis was conducted using machine learning (ML) approaches such as NB, KNN, Decision Tree (DT), and SVM models. The NB model fared better than the other algorithms, obtaining the highest F-score, accuracy, recall, and precision metrics. The study showed how SA can be used to better identify customer needs and enhance the quality of applications for mobile banking services. The objective of [10] was to use SA from Twitter to gather user satisfaction with digital banking in Indonesia. Information was collected for Indonesia's three digital banks: Blu, Jenius, and Jago. A total of 34,605 tweets were collected and examined. Naïve Bayes, K-nearest neighbors, logistic regression, support vector machines, decision trees, random forest, adaptive boosting, light Gradient Boosting, and extreme gradient boosting were the nine standalone classifiers used for SA. They employed two methods: hard voting and soft voting. With an F1-score of 73.34%, the work's findings show that the SVM performs the best when compared to other standalone classifiers. Soft voting

with the five best classifiers performed the best overall with an F1-score of 74.89%, and the ensemble technique outperformed the use of a stand-alone classifier. In [11], a dataset comprising reviews of governmental service applications was introduced. This dataset offers insights into the types of issues and requirements that users frequently report regarding these applications. It encompasses approximately 51,000 user reviews from six applications available on the Google Play Store. Data were obtained using web-scraping techniques. In [12], researchers employed sentiment analysis to aid the government in enhancing applications developed to combat the spread of COVID-19. They collected 8000 reviews from both the Google Play Store and App Store. Various ML algorithms, including Decision Tree, K-Nearest Neighbor, Support Vector Machine, and Naïve Bayes, have been utilized. The findings indicated that K-Nearest Neighbor yielded the most favorable outcomes, achieving an accuracy rate of 78.46%. An enhanced sentiment analysis technique was presented in [13] and used to evaluate user reviews of Saudi government applications that are accessible on the APP store and Google Play. This study used a new dataset of 51,000 user reviews. The Google pre-trained Word2Vec, BoWs, TF-IDF, AFINN, MPQA Subjectivity Lexicon, and Bing Liu lexicon were all incorporated as feature engineering techniques. The dataset was subjected to three ML models: Random Forest, Bagging, Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB). Using the SVM, the highest accuracy of 93.17% was achieved. The difficulties in sentiment analysis of Arabic text have been examined in [7], which also emphasizes how little is known about ASA in comparison to English and other Latin languages. Using a dataset of user comments from evaluations of certain mobile applications on Google Play Store, the study suggested a novel method for assessing sentiment in Arabic script. The strategy entailed enhancing data pretreatment algorithms, such as the Levenshtein distance (LD) algorithm, and merging it with the K-NN algorithm. The influence of effectively using the K-NN and LD algorithms for the ASA on mobile application reviews was examined through tests. According to the findings, the K-NN with the LD algorithm produced the best evaluation metrics for accuracy, recall, precision, and F-score. This study showed how the suggested method could increase the precision and potency of sentiment analysis in Arabic text. A study conducted in [14] sought to improve and sustain mobile banking applications while also learning what consumers thought of them. In this study, customer feedback gathered from mobile banking apps on Google Play Store was analyzed using application review-based sentiment analysis (SA). Three classes, positive, negative, and neutral, were manually assigned to the dataset. Arabic sentiment analysis was conducted using ML approaches such as NB, KNN, Decision Tree (DT), and SVM models. The NB model fared better than the other algorithms, obtaining the highest F-score, accuracy, recall, and precision metrics. The study showed how SA can be used to better identify customer needs and enhance the quality of applications for mobile banking services. A novel approach to ASA based on data from mobile app comments was presented in [15]. For data preprocessing, the study employed the LD method and discovered that when combined with the NB algorithm, the best accuracy of 96.40% was achieved for k = 9, whereas the NB algorithm achieved an accuracy of 95.80% for the same value of k. It is important to note that this study concentrated on Arabic sentiment analysis, and to the best of our knowledge, there is still a research gap in this area. The pre-trained Bidirectional Encoder Representations from Transformers (BERT) model was used in [16] to conduct sentiment analysis and topic modeling on social media messages to examine healthcare researchers' feelings toward ChatGPT. An early evaluation of ChatGPT's comprehension of thoughts, attitudes, and emotions within the text was conducted in [17]. Standard, open-domain, polarity shift, and sentiment inference evaluations were the four settings used in this assessment. To compare the performance of ChatGPT with that of the refined BERT and other cutting-edge models in end-task scenarios, they used five sentiment analysis tasks and 18 benchmark datasets. To learn more about ChatGPT's sentiment analysis skills, they also provided qualitative case studies and conducted human evaluations. In [18], researchers conducted sentiment analysis on Lyme illness using BERT and ChatGPT. Their research offers a useful manual for using Natural Language Processing (NLP) methods for sentiment analysis in the field of tick-borne illnesses. The researchers wanted to demonstrate how to assess the prevalence of bias in the discourse surrounding chronic manifestations of the disease. They demonstrated how to perform sentiment analysis using pre-trained language models using Python, utilizing a dataset of 5643 abstracts from scholarly journals on chronic Lyme disease. Using interpretable ML techniques and a novel methodology that uses cutting-edge large language models such as ChatGPT, the researchers validated their initial findings. To potentially lower the cost and complexity of such research, the study in [19] sought to determine whether ChatGPT could successfully substitute human-generated label annotations in social computing activities. Five significant datasets, including sentiment analysis, posture detection (twice), bot detection, and hate speech, were re-labeled using ChatGPT for evaluation. Although there are still certain obstacles to be addressed, their results indicate that ChatGPT has the capacity to manage these annotation jobs. With the best performance observed in the sentiment analysis dataset, ChatGPT accurately annotated 64.9% of the tweets, achieving an overall average accuracy of 0.609. However, researchers have observed notable differences in the performance between labels.

Future research examining the application of ChatGPT to human annotation tasks may draw inspiration from this study and use it as a starting point. This study demonstrates the primary motivation for examining the effects of ChatGPT on Arabic sentiment analysis. The study in [20] investigated the possibility of creating synthetic training data to support low-resource scenarios using data produced by stimulating a sizable generative language model, ChatGPT. They were able to outperform the already widely used methods for data augmentation by employing task-specific prompts for the ChatGPT. To evaluate and validate the quality of the created data, researchers have also investigated various approaches for assessing the similarity of the augmented data produced by ChatGPT. To address these problems, a straightforward yet powerful method for tweaking instructions was presented in [21]. Significant advancements in financial sentiment analysis have been accomplished by using this technique to refine a general-purpose Language Model and transform a small portion of supervised financial sentiment analysis data into instructional data. Their approach performed better than state-of-the-art supervised sentiment analysis models, including well-known Language Models such as ChatGPT and LLaMAs, particularly where contextual awareness and numerical comprehension are crucial. The experiment showed that their method performed better than other methods.

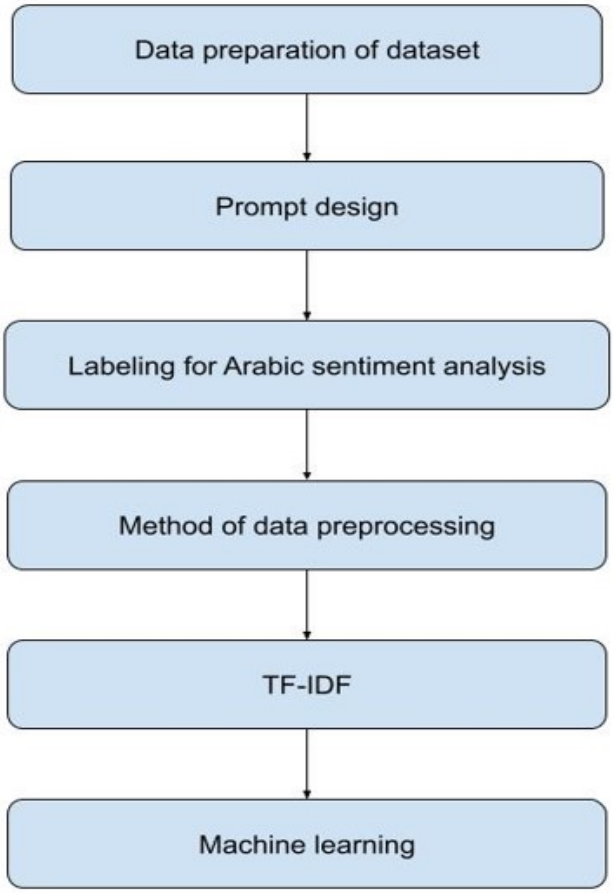

Figure 1. The proposed methodology's workflow

3. METHODOLOGY

This methodology is used to perform ASA on mobile application (app) reviews. The aim of this study was to investigate the impact of humans, ChatGPT, and Google Bard on ASA. The proposed approach for labeling ASA using humans, ChatGPT and Bard Google, involves two different approaches, which are outlined below. The approach consists of six main phases, as illustrated in Figure. 1.

3.1. Data preparation of dataset

This is the initial stage of the proposed approach. This process involves several steps. In the first step, we gathered "the evaluation data from the Google Play Store for several Yemeni mobile banking apps between March 5, 2021, and March 16, 2022. The data for eight distinct Yemeni financial applications were carefully gathered and extracted. Annajm, Railmobil, CAC Bank, Kuraimi Jawal, Hazmi Mobile, Moheet Mobile, Al Amal Bank, and Tadhamon Mobile are among the companies available on Google Play. Only aggregated reviews in Arabic were taken into consideration out of the 3197 reviews that were taken from Google Play. This database is currently available in Ref[1]. Table 1 shows the review data based on the apps used. 2940 Arabic-language user comments from mobile applications that have been classified as negative, neutral, or positive constitute the dataset used in this study. Figure 2 shows the frequency distribution of the polarity scores to help visualize their distribution in the training dataset. There were 1381 examples in the training dataset that were classified as positive remarks, 975 as negative comments, and 584 as neutral comments. Table 2 presents a summary of the different types of datasets used to evaluate the performance of the proposed model for ASA. Figures 3. Shows Words Cold.

Table 1. Statistics of the Arb-Apps comments dataset.

No	Applications Name	Comments Number
1	Annajm	328
2	Railmobile	669
3	Cac Bank	355
4	Kuraimi Jawal	727
5	Hazmi Mobile	114
6	Moheet Mobile	369
7	Al Amal Bank	304
8	Tadhamon Mobile	331
	Comments Total	3197

Other data were collected, as listed in Table 4. This

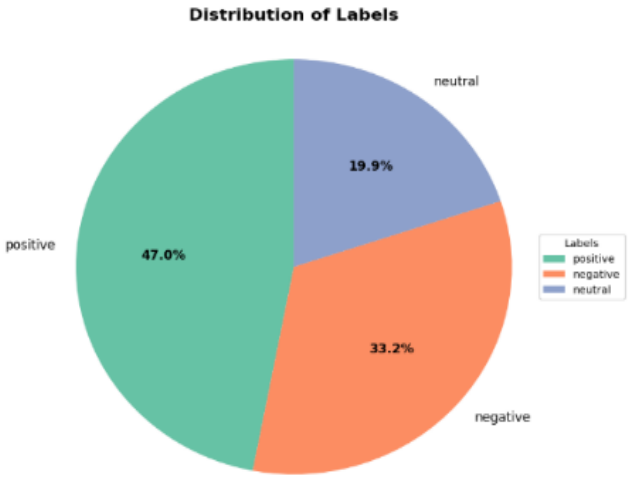

Figure 2. Distribution of the polarity scores for labeling using manually by humans.

Figure 3. Words Cold for ASA on dataset Mobile banking applications.

study aimed to investigate the impact of Google Bard on ASA. The study included 207 comments. The proposed method for labeling ASA using Bard Google involves two different approaches, which are outlined below. In this paper there are several stages in the initial stage, we collected “user comments” review data for several mobile banking applications in Yemen taken from the Google Play Store for the period from August **15**, 2023, to September **10**, 2023. The data was extracted and collected manually for five different Yemeni banking applications, it is available on Google Play, and they are (YKB Family, YCB, YKIB DIGITAL, ONECASHYE, and DAWLI MOBILE). Only aggregated Arabic reviews, the total number of reviews was **250**, extracted from Google Play, **and 207** were considered. Table 3 shows the statistics for the reviews based on their apps. Table 3 Shows statistics for reviews based on their apps.

3.2. Prompt design

The word "prompt" describes a set of guidelines used to design an LLM in the second phase, which aims to direct and improve its capability and purpose. Subsequent interactions with the model and its results can be influenced by a prompt. Therefore, to accomplish the intended result for a given job, it is essential to explicitly identify the appropriate prompts. There are three types of prompts: zero-, one-, and few-shot. We conducted a pilot experiment to identify the best prompt for our sentiment analysis task, which we now discuss [22, 23, 24, 25] (Kadaoui, K. et al., 2023). (B. Zhang et al., 2023). (S. Y. Kwon et al., 2023). Alto, V. (2023). DL: The act of locating and labeling unprocessed material (such text or pictures) with pertinent information so that artificial intelligence (AI) models can comprehend and learn from it.

Table 2. Datasets for ASA.

No	Dataset name	Description	Number of pairs
1	Mobile banking applications	Gathered "user comments" evaluation information from the Google Play Store for a number of Yemeni mobile banking apps between March 5, 2021, and March 16, 2022.	3197 user comment reviews
2	Mobile banking applications	We gathered "user comments" review information for a number of Yemeni mobile banking apps from the Google Play Store between August 15, 2023, and September 10, 2023. The information was carefully gathered and retrieved for five distinct banking applications in Yemen.	207 user comment reviews

Table 3. Statistics of the Arb-Apps Comments dataset

No	Applications Name	Comments Number
1	YKB Family	19
2	YCB	18
3	YKIB DIGITAL	10
4	ONECASHYE	44
5	Dawli mobile	103
Comments Total		207

Instructions: Prompt for ASA.

احتاج عرض **Sentiment analysis** للجمل التالية الذي عددها **20** فقرة بالترتيب بدون النص فقط تحديد احدى القطبيات **(Negative, Positive, Neutral):**

Input data: The results of the SA for the Arabic comments using the suggested approach (Assistant-Poe, Bing-Edge, Assistant-Poe with humans, Bing-Edge with humans, and Assistant-Poe with

Bing-Edge) are shown in Table 4, where the second column provides examples of comments. Each comment corresponds to the polarity result of each approach. It is frequently difficult to find high-quality DL for SA in particular domains or languages.

3.3. Data Labeling

The proposed six methods for classifying ASA utilizing humans, ChatGPT, and Bard Google are included in the third phase, which is the major body of the ASA and is subject to evaluation. In the subsequent subsection 3.3.2 labeling is done manually by humans, and in subsection 3.3.1, labeling issues are resolved with ChatGPT and Bard Google.

3.3.1. Data Labeling using manually by humans

This method entails human annotators manually annotating Arabic texts for SA. With 1381 positive records, the results indicate that the majority of the chosen comments were categorized as positive. There were 584 neutral records and 975 negative records. Thus, this method requires human annotators to manually annotate Arabic text for SA. After reading and evaluating the content, the annotators classified sentiments as good, negative, or neutral. This method acts as a standard to measure the accuracy of other labeling strategies. Table 5. Shows. Example of manual review data using humans. Figure 4. depicts the ASA word clouds. A word cloud or tag cloud is a type of data visualization for text data that illustrates the most frequently used terms and their intensities. The larger the term, the more frequently it is used, and smaller words have lower frequencies. The most commonly used word in the comments was "طبق," which translates to "App." Another widely used term is " خدم," which means "Servants." Arabic vocabulary such as "سدد'," "دخل"," "متز"," "شكل"," "عمل"," "جدت"," "حدث"," and " حول"are frequently used (Albahli, S. ,et al,2022).

Figure 4. Word cloud for use in comments that are positive (Right) or negative (Center) and neutral (lef) for data labeling by humans.

Furthermore, this method requires human annotators to manually categorize Arabic text for sentiment analysis. With 140 positive records, the results indicated that the majority of chosen comments were categorized as positive. There were twenty-three neutral records and

Table 4. Examples of comments and prompt for ASA.

No	Comment of reviews from mobile applications	Humans	Assistant-Poe	Bing-Edge	Assistant-Poe with humans	Bing-Edge with humans	Assistant-Poe with Bing-Edge
1	اخر تحديث احيانا يعلق مع تقديرنا للخدمات المضافة	Positive	Neutral	Negative	Negative	Neutral	Positive
2	ألأن ومع التحديث اصبح ظخم جداً جدا بس ثقيل	Positive	Negative	Neutral	Neutral	Negative	Positive
3	ممممممتاز بس كيف افتح البرنامج عبر صفحه الويب	Positive	Positive	Neutral	Positive	Positive	Positive
4	ياخي اشتي اسجل مارضيش يقلي لازم ان تكون	Negative	Negative	Neutral	Negative	Negative	Negative
5	نتمنى عمل خدمه عرض العمليات بتاريخ محدد أي يمكن التعرف على العمليات الذي تمت في أي وقت مثل تطبيق ام فلوس وليس فقط عشر عمليات	Neutral	Neutral	Positive	Neutral	Neutral	Neutral
6	من بعد تحديث التطبيق يفتح في الشاشة لون بنفسجي وتستمر طول الوقت ولا يفتح التطبيق ولا توصل لبياناته	Neutral	Negative	Negative	Negative	Negative	Negative

7	التطبيق بطيء جدا ضروري تشوفو حل عملتو لنا مشاكل مع الزبائن ساعه وحنا نحاول نسدد له رصيد وساعه وحنا نحاول نحولها الى باقة مزاياء لماذا لا تجعلوها كلها مره واحده مثل باقي التطبيقات	Negative	Negative	Neutral	Negative	Negative	Negative
8	التطبيق تمام بس في ملاحظات اتمنى يصلحوها الاولى انه ماتظهر شي الباقات اليومية النت الثانية الباقة الجديدة مزايا مكس ماتظهر ضمن العروض الثالثة فحص السلفة غير مفعلة وشكرا أبو عزام	Positive	Neutral	Neutral	Neutral	Neutral	Neutral
9	التطبيق كان ممتاز الان ولا بريال ارجوا اصلاحه تفعيل الباقات غير موجوده	Negative	Positive	Positive	Positive	Positive	Positive
10	التطبيق لا اعرف ماهي التحدثات مازال نفس المشكله لا تسديد باقات mtn ولايوجد تسديد سبا فون سوا تغير الواجه التطبيق لا اقل والاكثر	Negative	Negative	Negative	Negative	Negative	Negative
11	التطبيق لا باس فيه ولكن السوء فيه اعاده كلمة المرور لازم تروح لهم يعيدو ارسالها من الافضل يوجد اعاده كلمه المرور من التطبيق وارسالها على رقم الجوال	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
12	الله يبارك في جهودك... وخطوة جيدة نحو نجاح متجدد...مزيدا من التقدم والإزدهاااار...	Positive	Positive	Positive	Positive	Positive	Positive

Table 5. Example of review data using manually by humans

ID	Original Comments	Sentiment Polarity
1764	اروووع برنامج. جيد ممتاز بلتوفيق يانجم النجووم	Positive
2190	نريد نستخدم التطبيق في الريف كيف طريقة التفعيل	Neutral
28	اصبح سيء جدا حتى رقم الحواله الاكسبرس لايوجد في سجل العمليات	Negative
75	التحديث الاخير رووووووووووووعهاً جداً جد	Positive
168	التطبيق سيء جدا نرجو إصلاح مشكلة عدم فتح التطبيق	Negative
269	برنامج فاشل لم استطيع حتى فتحه من البدايه نرجو من القائمين عليه اصلاحه	Negative

forty-four negative records. Thus, this method requires human annotators to manually annotate Arabic text for sentiment analysis. After reading and evaluating the content, the annotators classified sentiments as good, negative, or neutral. This method acts as a standard to measure the accuracy of other labeling strategies. An example of review data is presented in Table 6.

Table 6. Example of review data.

ID	Original Comments	Sentiment Polarity
30	تطبيق رائع لافضل اسرة 😀انديلووووو اسرة بنك اليمن والكويت،.. احترامنا الكبير للقائمين على تطوير التطبيق.	Positive
65	تطبيق زباله ... ما توصل رساله التحقق للدخول الى البرنامج الا بعد ربع ساعة .واحيانا ما توصل..	Negative
125	برنامج جميل لمن لديه حساب في بنك اليمن الدولي	Positive
174	ممتاز جدا وفقكم الله الى المزيد من التطوير والتقدم ولو تجعلوا ارفاق الصور من الهاتف افضل من التقاط الصورة بشكل مباشر	Positive
106	أدخل بيانات الحساب عشان تفعيل الخدمة يقول لي: خطأ العميل غير موجود؟!!	Neutral
132	عتحدثووا التطبيق والا عتخلووه هكذا سااااع الزفت	Negative

3.3.2. Data Labeling using ChatGPT

An innovative method of ASA labeling with ChatGPT. In this paper, we suggest the following five methods for examining ASA labeling on ChatGPT:

- **Labeling using ChatGPT by Assistant-Poe**

This method involves labeling Arabic text for SA using the ChatGPT model (Assistant-Poe version). With 1329 positive records, the results indicate that the majority of the chosen comments were categorized as positive. While there were 371 neutral records, there were 1240 negative records. This method involves labeling Arabic text for SA using ChatGPT, a cutting-edge NLP model. Text can be categorized as having a positive, negative, or neutral sentiment using the Assistant-Poe version of ChatGPT. The frequency distribution of polarity scores in the training dataset is shown in Figure 5. Figure 6 shows the ASA word clouds.

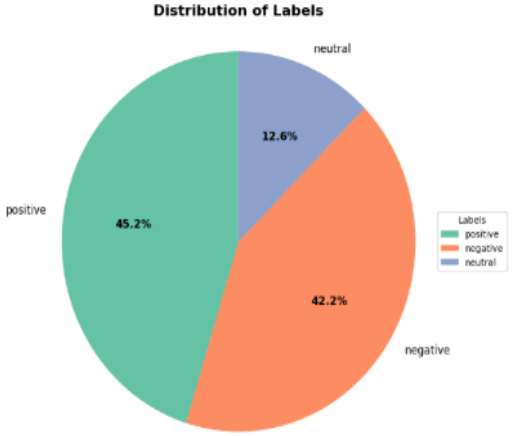

Figure 5. Distribution of the polarity scores for labeling using ChatGPT by Assistant-Poe.

Figure 6. Word cloud for use in comments that are positive (Right) or negative (Center) and neutral (lef) for data labeling using ChatGPT by Assistant-Poe.

- **Labeling using ChatGPT by Bing-Edge**

This method entails labeling Arabic text for SA using the ChatGPT model (Bing-Edge variation). With 1191 positive records, the results indicate that the majority of the chosen comments are categorized as positive. There were 973 records that were negative and 776 that

were neutral. This method involves labeling Arabic text for SA using ChatGPT, a cutting-edge NLP model. Text can be categorized as having good, negative, or neutral sentiments using ChatGPT's Bing-Edge version. The frequency distribution of polarity scores in the training dataset is shown in Figure 7. Figure 8 shows the ASA word clouds.

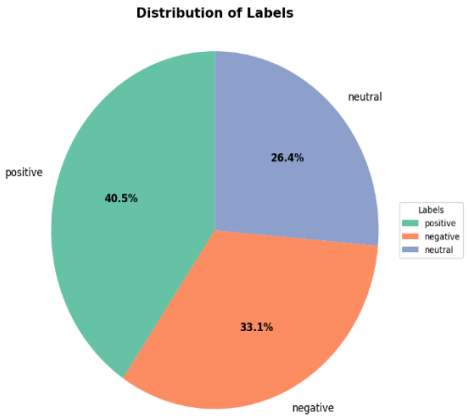

Figure 7. Distribution of the polarity scores for labeling using ChatGPT by Bing-Edge.

Figure 8. Word cloud for use in comments that are positive (Right) or negative (Center) and neutral (lef) for data labeling using ChatGPT by Bing-Edge.

- **Common labeling between Bing-Edge with humans**

This method entails labeling Arabic text for SA using the ChatGPT model (Bing-Edge with humans' variation). With 1339 positive records, the results indicate that the majority of the chosen comments were categorized as positive. While there were 475 neutral records, there were 1226 negative records. This method involves labeling Arabic text for SA using ChatGPT, a cutting-edge NLP model. Text can be categorized as having a positive, negative, or neutral emotion using Bing-Edge with the human version of ChatGPT. The frequency distribution of polarity scores in the training dataset is shown in Figure 9. Figure 10 shows the ASA word clouds.

- **Common labeling between Assistant-Poe with humans**

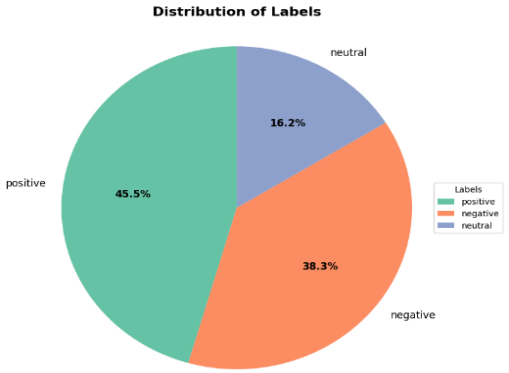

Figure 9. Distribution of the polarity scores for labeling using ChatGPT by Bing-Edge with humans.

Figure 10. Word cloud for use in comments that are positive (Right) or negative (Center) and neutral (lef) for data labeling using ChatGPT by Bing-Edge with humans.

This method involves labeling Arabic text for SA utilizing the ChatGPT model (Assistant-Poe with humans' variation). With 1340 positive records, the results indicate that the majority of the chosen comments are categorized as positive. There are 434 records that are neutral and 1166 records that are negative. This method entails labeling Arabic text for SA using ChatGPT, a cutting-edge NLP model. Text can be categorized as having a positive, negative, or neutral emotion using the Assistant-Poe with humans version of ChatGPT. The frequency distribution of polarity scores in the training dataset is shown in Figure 11. Figure 12 shows the ASA word clouds.

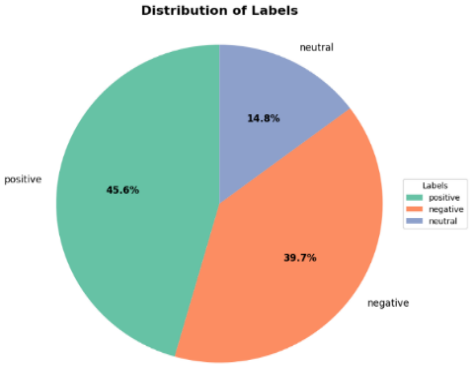

Figure 11. Distribution of the polarity scores for labeling using ChatGPT by Assistant-Poe with humans.

- **Common labeling between Assistant-Poe with Bing-Edge**

Figure 12. Word cloud for use in comments that are positive (Right) or negative (Center) and neutral (lef) for data labeling using ChatGPT by Assistant-Poe with humans.

This method involves labeling Arabic text for SA using the ChatGPT model (Assistant-Poe with Bing-Edge variation). With 1389 positive records, the results indicate that the majority of the chosen comments were categorized as positive. While there were 432 neutral records, there were 1119 negative records. This method involves labeling Arabic text for SA using ChatGPT, a cutting-edge NLP model. Text can be categorized as having a good, negative, or neutral sentiment using ChatGPT's Assistant Poe with the Bing-Edge version. The frequency distribution of polarity scores in the training dataset is shown in Figure 13. Figure 14 shows the ASA word clouds..

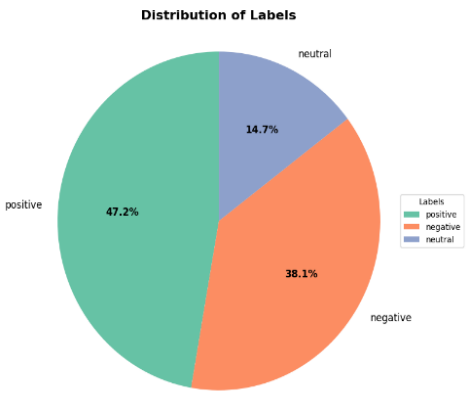

Figure 13. Distribution of the polarity scores for labeling using ChatGPT by Assistant-Poe with Bing-Edge.

Figure 14. Word cloud for use in comments that are positive (Right) or negative (Center) and neutral (lef) for data labeling using ChatGPT by Assistant-Poe with Bing-Edge.

The overall goal of the suggested method is to examine how well labeling techniques utilizing ChatGPT and people work for ASA. Using ChatGPT, a very sophisticated NLP model, offers the opportunity to increase the precision and effectiveness of ASA tagging.

- **Labeling using Bard Google**

A new method of ASA labeling using Bard Google. In this research, we suggest the following approach to examine ASA labeling on Bard Google. This method entails labeling Arabic text for sentiment analysis using Bard Google. With 139 positive records, the results indicated that the majority of chosen comments were categorized as positive. There were 13 neutral records and fifty-three negative records. This method entails labeling Arabic text for sentiment analysis using Bard Google, a cutting-edge natural language processing model. Texts can be categorized using Bard Google into three sentiment categories: neutral, negative, and positive. Overall, the proposed approach aims to investigate the effectiveness of labeling approaches using Google Bard for ASA. The use of Bard Google, a highly advanced natural language processing model, provides an opportunity to improve the accuracy and efficiency of labeling for ASA.

3.4. Data preprocessing

Case folding, filtering (removing punctuation), tokenization, stop word removal, stemming, and other text normalization techniques were used to preprocess the dataset in the fourth phase. a critical first step in improving and deriving valuable insights from data pre-processing. The accuracy of the analysis may be affected by flaws and inconsistencies in the data, which can be eliminated in this step. The following is a summary of the many methods used in data pre-processing [26] (Abbes, M., et al., 2023; Yahya, M. I., et al., 2022). The Preprocessing steps are shown in figure 15.

Data cleaning and homogenization were the goals of preprocessing. The data used in this study was text-based, therefore preprocessing was done on the data, which included case folding (replacing beginning Ġ, ġ, ĥ with), filtering (removing punctuation), tokenization, stopword removal, and stemming. The Khoja approach, which supports the Arabic language, was used in this study. It is anticipated that the data will become more consistent following preprocessing. The statistics of the preprocessed dataset are presented in Table 7.

3.4.1. Removal

The first step involves removing irrelevant or redundant data that do not contribute to the analysis. For example, the review contains the text
عندما اتصلت على"8001010بقيت منتظر الرد لاكثر من

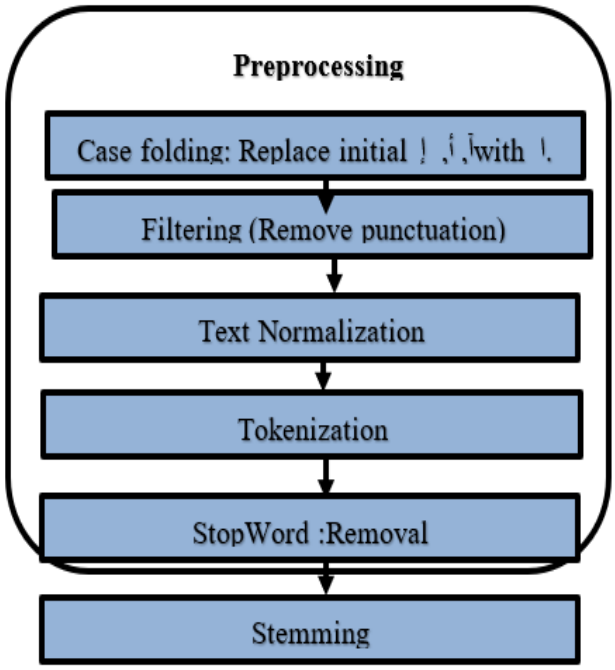

Figure 15. Preprocessing steps.

؟؟،جش يا لماذا دقائق10 " and the data "**8001010** ??" has been removed. The review contains the text "ممتاز تطبيق نجمات ٥ من اكثر يستحق يمكن لا" and the data "5" has been removed. Table 8. shows a sample of reviews on data removal.

Table 7. Dataset statistics has been preprocessed.

Unique word	**10377**
Not Letter	144
Punctuation	32
Stopword	489
Stemming word	8095
Not Stemming	1617

Table 8. A sample of reviews for the removal of data.

No	Sample of review	Removing
1	عندما اتصلت على8001010 بقيت منتظر الرد لاكثر من ١٠دقائق لماذا يا جش،؟؟	8001010,10 and ؟؟،
2	برنامج حلو بس متوقف حسابي فيه من ٣ ايام تقريبا. في صيانة او شي؟	3 and ?
3	برنامج تعبان عند تحويل باقات ابو ٤٥٠٠ توصل رصيد مابلا نتصل على ٣٣٣ يتم تفعيل الباقه	٤٥٠٠ and ٣٣٣
4	يستحق اكثر من ٥ نجمات تطبيق ممتاز	5

3.4.2. Folding of the case

In the second step, data were collected from uniform cases or letters contained in each comment from the mobile apps. Uniformizing letters was done from "آ، أ، إ" letters converted to "" letters and "ي ، ة"letters converted to "ى ، ه" letters , An example of replacing initial آ، أ، إ with ا.

3.4.3. Tokenizing

Tokenizing (splitting a text) is regarded as the third step, which is identified as the process of separating or cutting the input string based on each constituent word. The case-folding process for tokenization is marked by (-) as a list of word compilers. Tokenization involves splitting each string into a single word/token. In this process, each sentence or string is divided into several segments such as phrases, keywords, symbols, and words. These segments are known as tokens. Thus, the tags and punctuation marks were abandoned. Furthermore, the letters have been changed to lowercase. Table 9. shows an example of tokenizing a review sample.

Table 9. A sample of reviews for the tokening of data.

Sample of review	Tokenization
يستحق اكثر من ٥ نجمات تطبيق ممتاز	"يستحق" - "اكثر" - "من" - "نجمات" - "تطبيق" - "ممتاز"
مممممممممتاز بس كيف افتح البرنامج عبر صفحه الويب	"مممممممممتاز" - "بس" - "كيف" - "افتح" - "عبر" - "صفحه" - "الويب"

3.4.4. Stop word

The fourth step was filtering. Many terms (words) that are not important for identifying the polarity of a document are called Stop Words. The performance of sentiment analysis can be improved by reducing the size of the vector and eliminating these words. In addition, filtering is the process of removing unnecessary data from sentences and stop words. In this thesis, the Khoja method is used as it supports the Arabic language. Table 8. Shows a sample of reviews for the stop words in the data. Table 10. Show sample reviews for the stop word, which is not detected by Khoja, and the core of our work in this thesis is to improve the LD algorithm to delete these words, as shown in Table 11.

3.4.5. Rooting :Stemming

The last step stems from the use of words in a sentence where there are things that are not recognized by rules

or spelling dictionaries. Stemming is a common task in SA. In this study, we used the full stemmer to improve the performance of SA by reducing the size of the vector and finding the root of these words. Thus, the word vector size was considerably reduced. This study employed the Arabic stemmer offered by Khoja. Table 12. A sample of reviews on stemming is shown. Table 13. Sample reviews for stemming that are not detected by Khoja. The core of our work in this thesis is to improve the LD algorithm for detecting these words, as shown in Table 14. Thus, if a similar word is not found, the LD algorithm [7] is applied to obtain it from the root.

Table 10. A sample of reviews for the stop word of data.

No	Stop word	No	Stop word
1	أما	6	اول
2	الذين	7	به
3	لا	8	ولذلك
4	تحت	9	من
5	عليكم	10	في

Table 11. A sample of reviews for the stop word of data difficult to detect.

No	Stop word	No	Stop word
1	أأأأأأأاما	6	اااااااول
2	الذينااااااااااااااا	7	بببببببببببببه
3	لالالالا	8	ووووووووولذلك
4	تحتتتتتتت	9	منننننن
5	عليييييكم	10	ففففففففففففففي

Table 12. A sample reviews for stemming.

No	Original word	No	Original word
1	رووووووووووووعه	6	خدووووووم
2	ماقصرتو	7	برناااااج
3	مايشتغل	8	ررررررووووعههههه
4	الميلاد	9	ممتاااازة
5	بلغلط	10	عطططططططلان

3.5. Features Extraction

The fifth stage of the proposed method is the TF-IDF algorithm. The frequency with which a term appears in a document is measured using its Term Frequency (TF). This is the most straightforward way to evaluate a word's significance within a single page. The number of times a term "t" occurs in the document divided by the total number of terms in the document is the TF for that term. Different document lengths were taken into consideration in this standardization. As shown in Eq. (1).

$$\mathrm{TF}(t,d) = \frac{\text{Number of time t term appears in document d}}{\text{total number of terms in the document d}} \quad (1)$$

Inverse Document Frequency (IDF) evaluates a term's significance within a corpus or a collection of texts. It assists in determining the prevalence or rarity of a term across all publications. The logarithm of the total number of documents in the corpus divided by the number of documents containing the phrase indicates how the IDF for a term is calculated. Terms that appear frequently in papers are assigned less weight by this computation, making them less important. as Eq. (2).

$$\mathrm{IDF}(t,c) = \log\left(\frac{\text{total number of documents in corpus C}}{\text{total number of documents contining terms t}}\right) \quad (2)$$

The significance of words in the dataset was ascertained using this approach. It calculates each word's TF–IDF value according to how frequently it appears in the dataset documents. Words that appear in fewer papers have a higher TF-IDF value, whereas words that appear in more documents have a lower TF-IDF value. This method reduces the impact of frequent words that do not have any unique meaning, and assesses the importance of terms in expressing the text dataset (Zhao, C., et al., 2022; Durga, P., et al., 2023).

Table 13. A sample reviews for not stemming

ID	Comments of After Preprocessing	Sentiment Polarity
1764	اروووع برنامج. جيد مميز بالتوفيق نجم النجووم	Positive
2190	نرد نستخدم طبق ريف طرق فعل	Neutral
28	سيأ رقم حول الاكسبرس وجد سجل عمل	Negative
75	حدث خور رووووووووووووعه	Positive
168	التطبيق سيء رجأ صلح شكل عدم فتح التطبيق	Negative
269	برنامج فشل طوع فتح بدأ رجأ قوم صلح	Negative

Table 14. shows a sample of data that has been preprocessed.

3.6. Machine learning model

The sixth and final phase is ML, a branch of artificial intelligence that focuses on creating algorithms that allow computer systems to learn from data and gradually get better at what they do without explicit programming.

Table 14. A sample of data preprocessing.

No	Original word	Stemming	No	Original word	Stemming
1	انضباط	ضبط	6	بسيط	بسط
2	رائع	روع	7	الأصلي	صلي
3	جبار	جبر	8	ابداع	بدع
4	يحمله	حمل	9	خصوصا	خصص
5	عملي	عمل	10	ناجح	نجح

In SA, ML algorithms are used to determine the sentiment, whether neutral, positive, or negative, expressed in a particular text. The ML models that can be used for SA include NB, K-NN, SVM, and RF. Every algorithm has advantages and disadvantages, and the properties of the data being examined may affect how well or poorly it performs. In this study, we focus on NB, K-NN, SVM, and RF.

4. EXPERIMENTS AND RESULTS

This section presents the experiments carried out in this thesis to illustrate the proposed algorithms and compare them with different algorithms. The performance measurement is elaborated in sub-suction 3.4.1, sub-suction 3.4.2 and sub-suction 3.4.3. We will elaborate on a comparison study carried out to evaluate the performance of the proposed algorithms for ASA.

4.1. Evaluation criterion

To assess the efficacy of our proposed method, which entails labeling by hand by people, labeling by Assistant-Poe utilizing ChatGPT, Labeling with Bing-Edge's ChatGPT, we used four assessment metrics: Labeling using ChatGPT by Assistant-Poe with humans, Labeling using ChatGPT by Bing-Edge with humans, and Labeling using ChatGPT by Assistant-Poe with Bing-Edge. Accuracy, precision, recall, and F-score measurements were among these criteria. The ratio of correctly estimated samples to all anticipated samples is known as accuracy. In particular, it considers the number of true positives and true negatives in the data. We can obtain a more thorough understanding of our approach's performance and possible influence on the ASA field by employing a variety of evaluation metrics.

Accuracy: Accuracies for classification tasks, as in Eq. Eq. (3).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{3}$$

Recall: recall is computed using Eq. (4).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Precision: precision is computed using Eq. (5).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

F-score: F-score is computed using Eq. (6).

$$\text{F} - \text{score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{6}$$

TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively [18].

4.2. Experimental One for ML models

In this study, we used four ML techniques to SA Arabic text: NB, RF, SVM, and K-NN. We compared the accuracy of these techniques and applied them to six different methods of labeling ASA data: labeling by humans, labeling using ChatGPT by Assistant-Poe, labeling using ChatGPT by Bing-Edge, labeling using ChatGPT by Assistant-Poe with humans, labeling using ChatGPT by Bing-Edge with humans, and labeling using ChatGPT by Assistant-Poe with Bing-Edge. An accuracy scale was used to measure the performance of each method and technique, with higher accuracy indicating better performance. Four ML models were applied for the performance analysis: NB, RF, SVM, and K-NN. All models were fine-tuned to obtain the best results. Cross-validation was conducted to validate the performance of the applied models. We present the results of the experiment and compare the accuracy, precision, recall, and f-score measures of each technique and method, providing insights into the strengths and weaknesses of each approach for the SA of Arabic text. In figure 16. And figure 17. presents a confusion matrix that highlights the performance of ML models for the ASA. This matrix provides insight into the classification results for three sentiment categories: positive, negative, and neutral. The columns of the table represent the predicted values, whereas the rows represent the actual values, which presents a confusion matrix that highlights the performance of the ML models for the ASA. The evaluation results with the appropriate ML model for each measure are listed in Table 15. When compared with other K-NNs when k=2, the NB classifier performed better (accuracy, 89.65%; precision, 88.08%; recall, 88.43%; and F1-score, 88.25%) on the Google Play Store dataset of reviews of mobile banking applications in Yemen. The other SVM performed better (accuracy, 68.85%; precision, 56.08%; recall, 78.60%; and F1-score, 65.46%), while other DT techniques performed better (accuracy: 56.02%, precision: 43.22%, recall: 82.45%, and F1-score: 56.71%). The average values of classification accuracy, recall, precision, and F1-score of the NB, DT, KNN, and SVM techniques are graphically depicted in Figure 18. The accuracy results from the K-NN technique, specifically in the range of 87.89% to the value k = 2, while the accuracy results

from the combination of NB techniques, specifically in the range of 89.65%, perform better.

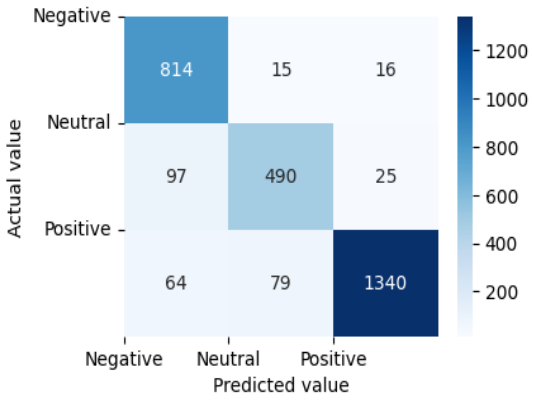

Figure 16. Confusion matrix of ASA using the NB algorithm by humans.

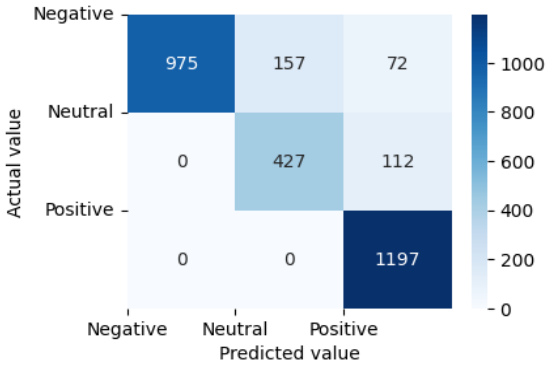

Figure 17. Confusion matrix of ASA using the K-NN algorithm by humans.

Table 15. The results ASA by the classifiers by humans.

ML	Accuracy	Recall	Precision	F-score
NB	**89.93%**	**88.14%**	**88.92%**	**88.53%**
K-NN	88.40%	86.60%	86.73%	86.66%
RF	47.01%	33.37%	49.00%	39.70%
SVM	70.54%	58.29%	79.77%	67.36%

Using manual labeling by humans, labeling using ChatGPT by Assistant-Poe, labeling using ChatGPT by Bing-Edge, labeling using ChatGPT by Assistant-Poe with humans, labeling using ChatGPT by Bing-Edge with humans, and labeling using ChatGPT by Assistant-Poe with Bing-Edge, we tested the efficacy of our suggested approach for ASA in this study. Four assessment metrics—accuracy, precision, recall, and F-score measures (Eqs 3-6)—formed the basis of the trials. By evaluating our method against other algorithms, we

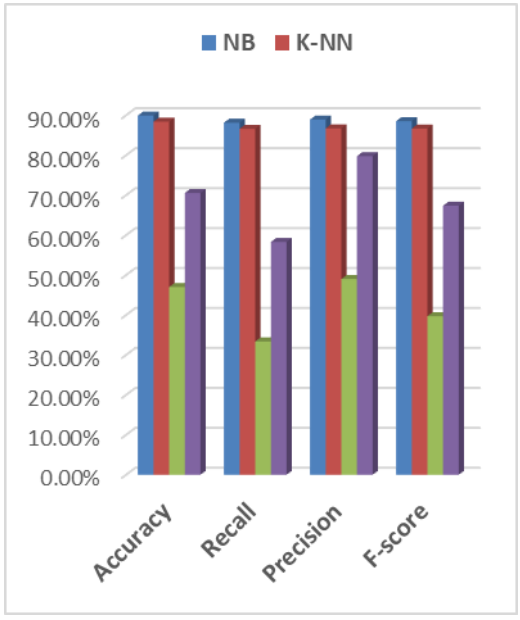

Figure 18. Values of classification accuracy, recall, precision and F1-score.

confirmed that it produces more accurate results. In particular, we outperformed the previous methods with an accuracy of 91.22%, precision of 89.62%, recall of 88.90%, and F-score of 89.26% for the suggested ChatGPT by Bing-Edge with people using the NB strategy. Table 16 summarizes these findings. Our results indicate that ASA can be enhanced by merging the Bing-Edge with the human version of the ChatGPT.

The outcomes of ASA utilizing K-NN are displayed in Table 17. Our method of using Bing-Edge's ChatGPT with people outperformed previous methods with an accuracy of 90.99%, precision of 87.38%, recall of 88.87%, and F-score of 88.12%. All things considered, our results indicate that enhancing ASA can be achieved by merging the Bing-Edge with the human version of ChatGPT.

Table 18 shows the results of the ASA using SVM. Our approach by Assistant Poe utilizing the SVM technique outperformed other approaches with an accuracy of 77.76%, precision of 52.51%, recall of 59.17%, and F-score of 55.64%. Overall, our results indicate that merging ChatGPT's Assistant Poe version is a useful strategy for raising the ASA.

The findings of the ASA using the RF method are presented in Table 19. Using the RF technique, our Assistant-Poe strategy outperformed other methods with an accuracy of 49.49%, precision of 45.54%, recall of 36.73%, and F-score of 40.66%. Overall, our results indicate that integrating ChatGPT's Assistant Poe version is a useful strategy for enhancing the ASA.

Table 16. The results ASA using NB technique..

Approach's	Accuracy	Recall	Precision	F-score
Manual labeling by humans	89.93%	88.14%	88.92%	88.53%
ChatGPT by Assistant-Poe	90.24%	88.00%	86.96%	87.48%
ChatGPT by Bing-Edge	88.23%	87.23%	88.47%	87.85%
ChatGPT by Assistant-Poe with humans	89.97%	88.54%	87.51%	88.02%
ChatGPT by Bing-Edge with humans.	**91.22%**	**89.62%**	**88.90%**	**89.26%**
ChatGPT by Assistant -Poe with Bing-Edge.	90.48%	88.82%	88.10%	88.46%

Table 17. The results ASA using K-NN technique.

Approach's	Accuracy	Recall	Precision	F-score
Manual labeling by humans	88.40%	86.60%	86.73%	86.66%
ChatGPT by Assistant-Poe	88.98%	75.62%	92.34%	83.15%
ChatGPT by Bing-Edge	85.68%	85.19%	86.08%	85.63%
ChatGPT by Assistant-Poe with humans	90.82%	87.71%	89.10%	88.40%
ChatGPT by Bing-Edge with humans.	**90.99%**	**87.38%**	**88.87%**	**88.12%**
ChatGPT by Assistant-Poe with Bing-Edge.	90.24%	87.00%	87.64%	87.32%

4.3. Experimental Two for ML models

In this study, we conducted an experiment to evaluate the effectiveness of the proposed approach for ASA using manual labeling by humans and labeling using Bard Google. The experiment was based on the following evaluation metric: accuracy. We validated the proposed approach by comparing its performance with those of other algorithms and found that it yielded more accurate results. Specifically, the proposed approach achieved an accuracy of 90.82%, outperforming other algorithms, such as K-NN with manual labeling by humans (accuracy of 87.44%). These results are summarized in Table 20 and depicted graphically in Figure 19., which indicate the average classification accuracy values. The proposed modifications to the algorithms resulted in a significant improvement in accuracy, with the proposed approach achieving the highest accuracy among all the algorithms tested.

Table 18. The results ASA using SVM technique.

Approach's	Accuracy	Recall	Precision	F-score
Manual labeling by humans	70.54%	58.29%	79.77%	67.36%
ChatGPT by Assistant-Poe	**77.76%**	**59.17%**	**52.51%**	**55.64%**
ChatGPT by Bing-Edge	69.97%	65.48%	73.87%	69.42%
ChatGPT by Assistant-Poe with humans	74.80%	59.22%	84.03%	69.48%
ChatGPT by Bing-Edge with humans.	75.51%	58.76%	84.39%	69.28%
ChatGPT by Assistant-Poe with Bing-Edge.	75.61%	58.54%	84.79%	69.26%

5. CONCLUSION

This study is dedicated to analyzing the Arabic sentiments of manually collected Arabic datasets. This is related to the mobile banking application users in Yemen. The sentiment analysis tasks were based on reviews collected from Google Play Store. The ML models used were naïve Bayes, K-nearest neighbor, and support vector machines. Thus, the advantages and disadvantages of a particular mobile application can be observed based on user reviews. The mobile banking of Yemen application providers can concentrate on correcting defects and better adjust to the demands of their users. These methods include manual labeling by humans and labeling using Bard Google. In this study,

Table 19. The results ASA using RF technique.

Approach's	Accuracy	Recall	Precision	F-score
Manual labeling by humans	47.01%	33.37%	49.00%	39.70%
ChatGPT by Assistant-Poe	**49.49%**	**36.73%**	**45.54%**	**40.66%**
ChatGPT by Bing-Edge	40.58%	33.41%	27.81%	30.35%
ChatGPT by Assistant-Poe with humans	46.29%	34.00%	43.08%	38.01%
ChatGPT by Bing-Edge with humans.	47.41%	34.88%	46.09%	39.71%
ChatGPT by Assistant -Poe with Bing-Edge.	48.95%	34.83%	45.93%	39.62%

Table 20. The results ASA using manual labeling by humans and labeling using Bard Google.

Approach's	Labeling using manually by humans	Labeling using Bard Google
K-NN	87.44%	90.82%
NB	98.07%	98.07%
SVM	90.82%	92.27%

we examined the effects of ChatGPT variations on ASA. We evaluated the effectiveness of six distinct approaches to data labeling for ASA using four ML techniques: NB, K-NN, SVM, and RF. These techniques include labeling by hand by humans, labeling by Bing-Edge, labeling by ChatGPT by Assistant-Poe, labeling by Bing-Edge, labeling by ChatGPT by Assistant-Poe with humans, labeling by Bing-Edge with humans, and labeling by Assistant-Poe with Bing-Edge simultaneously. Using different Bing-Edge models for ASA, our experimental results demonstrated that the NB approach performed the best, with an accuracy of 91.22%, recall of 89.62%, precision of 88.90%, and F-score of 89.26%. Furthermore, the findings imply that employing several models, especially Assistant-Poe models, can yield a greater accuracy than either a human-labeled dataset or a single-language model. Our results indicate that the NB technique Assistant-Poe

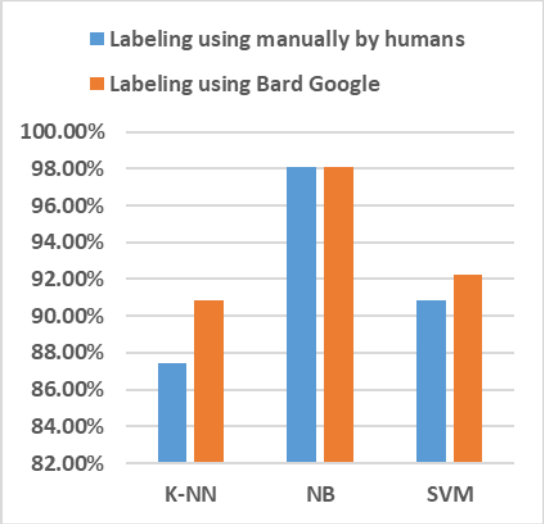

Figure 19. A graphical representation of the performance metrics for manual labeling by humans and labeling using Bard Google.

model is a useful strategy for ASA. Future research should increase the amount of data (Big Data) and include reviews in both Arabic and English. It may also look into other ML techniques or language models for sentiment analysis as well as the effects of various labeling strategies on the precision of sentiment analysis in Arabic text.

REFERENCES

[1] M. E. Permana et al. "Sentiment Analysis and Topic Detection of Mobile Banking Application Review". In: *Fifth International Conference on Informatics and Computing (ICIC)*. IEEE, 2020, pp. 1–6.

[2] Y. Bidulya and E. Brunova. "Sentiment analysis for bank service quality: a rule-based classifier". In: *IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, 2016, pp. 1–4.

[3] E. Brunova and Y. Bidulya. "Aspect extraction and sentiment analysis in user reviews in Russian about bank service quality". In: *11th IEEE International Conference on Application of Information and Communication Technologies (AICT)*. 2019, pp. 1–4. DOI: 10.1109/ICAICT.2017.8687070.

[4] F. Alqasemi et al. "Arabic Poetry Meter Categorization Using Machine Learning Based on Customized Feature Extraction". In: *International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE)*. IEEE, 2021, pp. 1–4.

[5] Bramanthyo Andrian et al. "Sentiment Analysis on Customer Satisfaction of Digital Banking in Indonesia". In: *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* 13.3 (2022).

[6] M. Al-Shamani et al. "Designing an Arabic Google Play Store User Review Dataset for Detecting App Requirement Issues". In: *Advances on Smart and Soft Computing*. Springer, Singapore, 2022, pp. 133–143.

[7] A. A. Al-Shalabi et al. "Investigating the Impact of Utilizing the K-Nearest Neighbor and Levenshtein Distance Algorithms for Arabic Sentiment Analysis on Mobile Applications". In: *JAST* 1.2 (2023).

[8] K. Kadaoui et al. "TARJAMAT: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties". In: *arXiv preprint arXiv:2308.03051* (2023).

[9] P. P. Ray. “ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope”. In: *Internet Things Cyber-Physical Syst.* (2023).

[10] S. Al-Hagree and G. Al-Gaphari. “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)*. IEEE, 2022, pp. 1–4.

[11] S. Al-Hagree and G. Al-Gaphari. “Arabic Sentiment Analysis on Mobile Applications Using Levenshtein Distance Algorithm and Naive Bayes”. In: *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*. IEEE, 2022, pp. 1–6.

[12] Bramanthyo Andrian et al. “Sentiment Analysis on Customer Satisfaction of Digital Banking in Indonesia”. In: *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* 13.3 (2022).

[13] M. Hadwan et al. “Arabic Sentiment Analysis of Users’ Opinions of Governmental Mobile Applications”. In: *Comput. Mater. Continua* 72.3 (2022), pp. 4675–4689.

[14] S. Al-Hagree and G. Al-Gaphari. “Arabic Sentiment Analysis Based Machine Learning for Measuring User Satisfaction with Banking Services’ Mobile Applications:

[22] B. Zhang, H. Yang, and X. Y. Liu. “Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models”. In: *arXiv preprint arXiv:2306.12659* (2023).

[23] S. Y. Kwon et al. “ChatGPT for Arabic Grammatical Error Correction”. In: *arXiv preprint arXiv:2308.04492* (2023).

[24] V. Alto. *Modern Generative AI with ChatGPT and OpenAI Models*. UK: Packet Publishing. O’Reilly, 2023.

[25] M. Abbes, Z. Kechaou, and A. M. Alimi. “A Novel Hybrid Model Based on CNN and Bi-LSTM for Arabic Multi-domain Sentiment Analysis”. In: *Conference on Complex, Intelligent, and Software Intensive Systems*. Cham: Springer Nature Switzerland, 2023, pp. 92–102.

Comparative Study”. In: *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*. IEEE, 2022, pp. 1–4.

[15] S. Al-Hagree and G. Al-Gaphari. “Arabic Sentiment Analysis on Mobile Applications Using Levenshtein Distance Algorithm and Naive Bayes”. In: *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*. IEEE, 2022, pp. 1–6.

[16] S. V. Praveen and V. Vajrobol. “Understanding the perceptions of healthcare researchers regarding ChatGPT: a study based on bidirectional encoder representation from transformers (BERT) sentiment analysis and topic modeling”. In: *Ann. Biomed. Eng.* (2023), pp. 1–3.

[17] Z. Wang et al. “Is ChatGPT a good sentiment analyzer? A preliminary study”. In: *arXiv preprint arXiv:2304.04339* (2023).

[18] T. Susnjak. “Applying BERT and ChatGPT for sentiment analysis of Lyme disease in scientific literature”. In: *arXiv preprint arXiv:2302.06474* (2023).

[19] Y. Zhu et al. “Can ChatGPT reproduce human-generated labels? A study of social computing tasks”. In: *arXiv preprint arXiv:2304.10145* (2023).

[20] S. Ubani, S. O. Polat, and R. Nielsen. “Zero Shot Data Aug: Generating and Augmenting Training Data with ChatGPT”. In: *arXiv preprint arXiv:2304.14334* (2023).

[21] K. Kadaoui et al. “Tarjamat: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties”. In: *arXiv preprint arXiv:2308.03051* (2023).

[26] M. I. Yahya et al. “Spelling Correction Using the Levenshtein Distance and Nazief and Adriani Algorithm for Keyword Search Process Indonesian Qur’an Translation”. In: *2022 Seventh International Conference on Informatics and Computing (ICIC)*. IEEE, 2022, pp. 01–06.