مجلــة العلــوم
الهندسية والتقنية
مجلة علمية محكمة - تصدر عن كلية الهندسة وكلية الحاسبات والمعلوماتية – جامعة ذمار

Using Machine Learning for Arabic Sentiment Analysis in Higher Education: Investigating the Impact of Utilizing the ChatGPT and Bard Google

Salah AL-Hagree[1,2], Ghaleb Al-Gaphari[1], Fuaad Hasan Abdulrazzak[3], Maher Al-Sanabani[3], Ahmed Al-Shalabi[1]

[1] Computer Science Department, Faculty of Computer and Information Technology, Sana'a University, Yemen
[2] Computer Science Department, Faculty of Sciences, Ibb University, Yemen.
[3] Computer Science Department, Faculty of Computer and Information Technology, Thamar University, Yemen.
s.alhagree@gmail.com, s.alhagree@su.edu.ye, drghalebh@su.edu.ye, fuaad.abdulrazzak@tu.edu.ye, M.sanabani@gmail.com , a.alshalabi@su.edu.ye

Abstract.

Gaining an understanding of the application's quality and meeting, the user's needs are crucial in the development of applications. To enhance the quality of applications, it is important to comprehend the requirements of the users. One effective approach for achieving this is through the utilization of application review-based sentiment analysis (SA). In this study, the objective was to assess students' opinions regarding mobile applications of universities in order to update and maintain them accordingly. Mobile applications of universities have become an integral part of students' lives, thus making it imperative to analyze user comments on these apps for SA purposes, where student input is crucial for assessing the effectiveness of educational institutions. This paper presents a machine learning (ML) based approach to sentiment analysis on students' evaluation of higher education institutions. The study analyzes a corpus containing approximately 275 student reviews written in Arabic, It also evaluates the performance of three ML techniques, including K-Nearest Neighbors (K-NN), Decision Tree (DT), and Random Forest (RF) using an accuracy, precision, recall, and f-score measures. In addition, the study compares one method of labeling the data for ASA, including manual labeling by humans, labeling by Bard Google and labeling by ChatGPT. Experimental results show that the K-NN technique performed the best, achieving an accuracy of 74.91% by ChatGPT models for Arabic sentiment analysis (ASA). Moreover, utilizing proposed active labeling method with Bard Google achieved higher accuracy compared to other labeling methods. The study proposed study suggests that the K-NN technique with ChatGPT models and proposed active labeling method are effective approaches for ASA by ChatGPT. It is indicated by the empirical results that promising results are yielded on the evaluation of students' opinions of higher educational institutions by an ML based approach.

Keywords: *Arabic Sentiment Analysis, Higher Educational, Bard Google, ChatGPT, Machine learning.*

I. Introduction

Over the past few years, there has been a growing fascination with sentiment analysis (SA) within the field of text mining. It has gained popularity as a prominent research area in higher education, specifically in the realm of opinion mining. This field focuses on analyzing and comprehending students' opinions regarding their educational institutions, aiming to enhance the decision-making process by improving its overall quality. This study aimed to investigate the utilization of SA in higher education by conducting a comprehensive analysis. It sought to identify and categorize the commonly employed and effective SA techniques and methods within the higher education domain. SA involves employing natural language processing techniques, text analysis, and statistical methods to analyze subjective information, such as opinions, attitudes, impressions, and emotions. Its purpose is to extract and categorize students' opinions and emotions, which are crucial in various aspects [1]. This approach finds wide application in processing, searching, and extracting factual information from diverse platforms such as blogs, Google Play , Tumblr, Instagram, Twitter, Facebook, and more [2]. As students form a significant part of universities, their perspectives and opinions play a vital role in enhancing teaching and addressing institutional challenges and concerns. Therefore, evaluating student opinions in digital educational resources and learning environments becomes essential for assessing the institution, teachers, and teaching effectiveness. Smartphone applications have become increasingly integral to our daily lives, experiencing a surge in usage. The Google Play Store is a widely recognized platform that provides access to a diverse range of Android applications. Among these applications, those related to higher education play a crucial role in enhancing the delivery of services to students with greater efficiency and effectiveness [3]. In SA 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 conducting SA: lexicon-based sentiment analysis (LBSA) and ML Arabic sentiment analysis (MLSA). LBSA utilizes a vocabulary dictionary to calculate the polarity of each text record, while 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 prior to the prediction process [4]. SA encounters a challenge when it comes to capturing the experiences of university students in higher education. Presently, numerous companies have emerged with the advent of artificial intelligence, offering a range of tools designed to assist students in their higher education journey. Among the most renowned tools in this domain are those powered by artificial intelligence, such as

Lectomate[1], Qonqur[2], Quetab [3], Dunno [4], Conker [5], Tutor AI [6], Resoomer[7], Lectomate [8], Gistvid [9] [10]. The Arabic language holds significant popularity and widespread usage, making it crucial to employ SA tools for Arabic adoption. However, the complexity of the Arabic language, including its morphology, structure, and variations, poses challenges. Further efforts are needed to improve Arabic language SA [1]. In this study, a specific Arabic dataset was manually collected and intentionally annotated SA tasks. The K-Nearest Neighbors (K-NN), Decision Tree (DT), and Random Forest (RF) algorithms were utilized to conduct ASA. The study focused on the challenges faced in SA within higher education institutions, with a specific focus on mobile applications in Yemeni universities. The models were evaluated using accuracy metrics for ML evaluation, and a comparison among the three models revealed the superiority of the K-NN algorithm with ChatGPT.

The structure of this paper is as follows. Section 2 provides an overview of recent studies on SA that utilize app feedback data from Google Play, 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. Lastly, Section 5 presents the conclusions of the study.

II. LITERATURE REVIEW

To identify the research gap in SA, we conducted a thorough review of important and relevant studies. SA is the process of extracting patterns from textual data, which can include categorizing and interpreting sentiment into negative, positive, or neutral comments using techniques such as ML. With the increasing availability of user information on the web, including social networks and other platforms, SA has become an important tool for understanding user opinions and behaviors. Many studies have focused on developing solutions for SA based on ChatGPT and Bard Google. Additionally, reviews on SA are often based on various

[1] https://topai.tools/t/bloombot-ai
[2] https://topai.tools/t/qonqur /
[3] https://topai.tools/t/quetab
[4] https://topai.tools/t/Dunno
[5] https://topai.tools/t/conker
[6] https://topai.tools/t/tutor-ai
[7] https://topai.tools/t/resoomer-com
[8] https://topai.tools/t/lectomate
[9] https://topai.tools/t/Gistvid

platforms, such as Google Play. By exploring these studies and platforms, we aim to contribute to the field of ASA and provide a more effective and accurate approach to analyzing sentiment in Arabic text. In [1] the previous study explored the challenges of SA in Arabic language and the lack of research on ASA compared to English and other Latin languages. The study proposed a new approach to analyzing sentiment in Arabic script using the comments dataset of users of some mobile applications reviews available on the Google Play Store. The approach involved improving algorithms such as the Levenshtein distance (LD) algorithm for data preprocessing and combining it with the K-NN algorithm. The study conducted experiments to investigate the impact of utilizing the K-NN and LD algorithms for ASA on mobile applications reviews effectively. The results showed that the K-NN with LD algorithm achieved the highest accuracy, recall, precision, and F-score evaluation measures. The study demonstrated the potential of the proposed approach to improve the accuracy and effectiveness of SA in Arabic text. In [2], the previous study aimed to understand customer opinions towards mobile banking services' applications and to improve and maintain these applications. The study used application review-based SA to analyze user comments collected from banking mobile apps on Google Play Store. The dataset was labeled manually into three classes: positive, negative, and neutral. ML techniques, including NB, KNN, DT, and SVM models, were utilized for ASA. The NB model outperformed the other algorithms, achieving the highest accuracy, recall, precision, and F-score measures. The study demonstrated the potential of using SA to understand user requirements and improve application quality in mobile banking services. In[4], the study introduces a machine learning (ML) based method for SA, focusing on students' evaluations of higher educational institutions. The researchers analyze a dataset consisting of approximately 700 student reviews written in Turkish. They employ conventional text representation schemes and ML classifiers in their analysis. In the experimental analysis, three conventional text representation schemes (term-presence, term-frequency, and TF-IDF) and three N-gram models (1-gram, 2-gram, and 3-gram) are considered alongside four classifiers (support vector machines, Naïve Bayes, logistic regression, and RF algorithm). Additionally, the predictive performance of four ensemble learners (AdaBoost, Bagging, Random Subspace, and voting algorithm) is evaluated. The empirical findings demonstrate that the ML based approach shows promising results in assessing students' evaluations of higher educational institutions. In [5], the study aimed to systematically review the recent advancements in the application of SA in higher education. The primary objectives were to categorize the commonly used and successful SA techniques and methods within the higher education domain. The researchers conducted a systematic mapping review of 840 articles, ultimately selecting 22 relevant studies based on predetermined criteria. The

findings indicated that the selected studies primarily focused on six domains in applying SA within higher education, with a particular emphasis on evaluating teaching quality. The study also revealed that the utilization of specific SA techniques could prove to be a valuable tool for institutions in addressing specific learning challenges, improving the quality of higher education institutions, and evaluating the teaching process and teachers' performance. In [6], to analyze healthcare researchers' emotions towards ChatGPT, the researchers utilized the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model and conducted SA and topic modeling on social media posts. In [7], the researchers conducted an initial assessment of ChatGPT's ability to comprehend opinions, sentiments, and emotions within the text. This evaluation included four settings: standard evaluation, open-domain evaluation, polarity shift evaluation, and sentiment inference evaluation. To carry out this evaluation, they utilized 18 benchmark datasets and 5 SA tasks, comparing ChatGPT, performance with fine-tuned BERT and other state-of-the-art models in end-task scenarios. Additionally, they conducted human evaluation and presented qualitative case studies to gain further insight into ChatGPT's SA capabilities. In [8], the researchers utilized BERT and ChatGPT to perform SA on Lyme disease. Their study provides a practical guide to conducting SA in the domain of tick-borne diseases using Natural Language Processing (NLP) techniques. The researchers aimed to demonstrate how the occurrence of bias in the discourse neighboring chronic manifestations of the disease can be evaluated. They used a dataset of 5643 abstracts from academic journals on the topic of chronic Lyme disease to show the steps involved in conducting SA using pre-trained language models with Python. The researchers validated their preliminary results using interpretable ML tools and a novel methodology that employs emerging state-of-the-art large language models like ChatGPT. In [9], the researchers aimed to investigate whether ChatGPT could effectively replace human-generated label annotations in social computing tasks, potentially reducing the cost and complexity of such research. To test this, they used ChatGPT to re-label five influential datasets covering SA, stance detection (twice), bot detection, and hate speech. Their findings suggest that ChatGPT does have the potential to handle these annotation tasks, although there are still some challenges to overcome. Overall, ChatGPT achieved an average accuracy of 0.609, with the highest performance observed in the SA dataset, correctly annotating 64.9% of tweets. However, the researchers noted significant variation in performance across different labels. This study has the potential to inspire new avenues of analysis and serve as a foundation for future research exploring the use of ChatGPT for human annotation tasks. However, the present study showed us a main motive to investigate the impact of ChatGPT on the analysis of sentiment of the Arabic language.

III. Methodology

This study aimed to investigate the impact of ChatGPT and Bard Google on ASA. It included 275 comments. The proposed method for labeling ASA using ChatGPT and Bard Google involves two different approaches, which are outlined below. The approach consists of five main phases, as illustrated in Fig. 1. In this paper, there are several stages. In the initial stage, the researchers collected "student comments" and review data for several mobile applications in Yemeni universities 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 six different Yemeni universities applications, which are available on Google Play. The list includes (Sana'a University, Thamar University, University of Science and Technology, Yemeni Jordanian University, Queen Arwa University, and *Unified Electronic Coordination Portal for Yemeni Universities*.

A. Data collection

This is the first phase of the approach. It included only aggregated Arabic reviews. The total number of reviews was 300 extracts from Google Play, 275 were considered. Table 1 shows statistics for reviews based on their apps.

B. Labelingmanually by humans, Labeling by ChatGPT, and Labeling by Bard Google.

The second phase was divided into three approaches as follows:

- **Labeling manually by humans**

This approach involves manually labeling Arabic text for SA by human annotators. The results show that the majority of the selected comments are classified as positive, with 150 positive records. The number of negative records is 102, while the number of neutral records is 23. Therefore, this approach involves labeling Arabic text for SA manually by human annotators. The annotators read and analyze the text to determine the sentiment expressed and label it as positive, negative, or neutral. This approach serves as a benchmark for evaluating the accuracy of the other labeling approaches.

- **Labeling by ChatGPT**

A novel approach to labeling for ASA utilizing ChatGPT. In this paper, we propose one method for investigating labeling of ASA using ChatGPT as follows:

This approach involves using ChatGPT to label Arabic text for SA. The results show that the majority of the selected comments are classified as positive, with 156 positive records. The number of negative records is 99, while the number of neutral records is 20. This approach involves using ChatGPT, a state-of-the-art natural language processing model, to label Arabic text for SA. ChatGPT is used to classify text into positive, negative,

or neutral sentiment categories. Overall, the proposed approach aims to investigate the effectiveness of labeling approaches by ChatGPT for ASA. The use of ChatGPT, a highly advanced natural language processing model, provides an opportunity to improve the accuracy and efficiency of labeling for ASA.

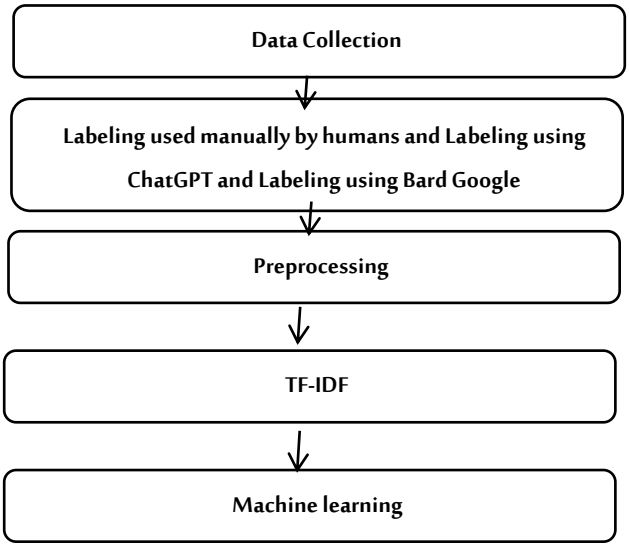

Fig. 1. Data collection and preprocessing steps.

TABLE 1. STATISTICS OF THE ARB-APPS COMMENTS DATASET.

No	Applications Name	Arabic Name	Comments Number
1	Sana'a University [10]	جامعة صنعاء	72
2	Thamar University [11]	جامعة ذمار	127
3	UST-DEV-TEAM [12]	جامعة العلوم والتكنولوجيا	18
4	ALSHIBANI Group [13]	الجامعة اليمنية الأردنية	3
5	QAU Dev. [14]	جامعة الملكة أروى	42
6	YCIT-HE. [15]	بوابة التنسيق الأهلي	13
Comments Total			**275**

10 https://play.google.com/store/apps/details?id=libosft.ye.com.sanaunif2

11 https://play.google.com/store/apps/details?id=thamar.univ.studentgate

12 https://play.google.com/store/apps/details?id=com.saqib.uststudentapp

13 https://play.google.com/store/apps/details?id=com.it.group

14 https://play.google.com/store/apps/details?id=edu.qau.queenarwauniversity.yemen

15 https://play.google.com/store/apps/details?id=org.ycit_he.p.nasseq

- **Labeling by Bard Google**

This is a novel approach to labeling for ASA utilizing Bard Google. In this paper, we propose one methods for investigating labeling of ASA using on Bard Google as follows: This approach involves using the Bard Google to label Arabic text for SA. The results show that the majority of the selected comments are classified as positive, with 135 positive records. The number of negative records is 89, while the number of neutral records is 51. This approach involves using Bard Google, a state-of-the-art natural language processing model, to label Arabic text for SA. Bard Google is used to classify text into positive, negative, or neutral sentiment categories. Overall, the proposed approach aims to investigate the effectiveness of labeling approaches by Bard Google 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.

C. Pre-processing

In the third phase, the dataset is preprocessed by removing stop words, stemming, and applying other text normalization techniques. Data pre-processing is a crucial step in enhancing and extracting meaningful insights from data. This step helps remove inconsistencies and errors that may be present in the data, which can affect the accuracy of the analysis. There are several techniques involved in data pre-processing, which are summarized as follows [1]: Table 2. shows example of review data and Table 3 shows dataset statistics that have been preprocessed.

- **Removal**: This involves removing any irrelevant or redundant data that does not contribute to the analysis.
- **Folding of Case**: This step involves converting all the text data to a standard case; typically lowercase, to reduce the number of unique words in the dataset.
- **Tokenization**: This step involves breaking down the text data into individual words, known as tokens, to facilitate analysis.
- **Stop Word Filtering**: This technique involves removing commonly used words, known as stop words, from the dataset, as they do not add any significant value to the analysis.
- **Rooting or Stemming**: This step involves reducing the words in the dataset to their root form, which helps to reduce the number of unique words and improve the accuracy of the analysis.

TABLE 2. EXAMPLE OF REVIEW DATA.

ID	Original Comments	Sentiment Polarity
12	تطبيق في قمه الروعه والفكره ممتازه جدا مماخلقت لدينا حجات كثيره جدا .....واهمها ارتباط الطالب بالجامعه وووو.	Positive
32	لا أرى فيه أي فائده للطالب الجامعي في جامعة صنعاء	Negative
250	الف شكر لجامعة ذمار على هذا الانجاز وكذلك ملتقى الطالب الجامعي	Positive
264	قوه القوه هذا التنسيق عن بعد يتيح للطلاب التسجيل من اي مكان وفي لي وقت	Positive
186	ارجوا منكم اتاحة تعديل البيانات الشخصية!!	Neutral
98	البرنامج مايفتح تظهر شعار الجامعه فقط ولايفتح شيئ	Negative

TABLE 3. DATASET STATISTICS WHICH HAVE BEEN PREPROCESSED.

Unique word	2326
Not Letter	19
Punctuation	32
Stopword	473
Stemming word	1586
Not Stemming	216

D. Features Extraction:

TF-IDF algorithm is the fourth phase of the proposed approach. This algorithm is used to determine the importance of words in the dataset. It calculates the TF-IDF value for each word based on its frequency in the documents within the dataset. The TF-IDF value is increased for words that appear in fewer documents and decreased for words that appear in more documents. This technique is used to determine the significance of words in representing the text dataset and to reduce the impact of common words that do not provide any distinctive meaning [1].

E. Machine learning (ML)

In the fifth phase, and the last ML is a subfield of artificial intelligence that involves designing algorithms that enable computer systems to learn from data and improve their performance over time without being explicitly programmed. In SA and ML, algorithms are used to identify the sentiment expressed in a given text,

such as whether it is negative, positive, or neutral. K-NN, DT, and RF, are all ML algorithms that can be used for SA. Each algorithm has its strengths and weaknesses and may perform better or worse depending on the characteristics of the data being analyzed.

- **K-Nearest Neighbors (K-NN)**

K-NN is a non-parametric algorithm that classifies a given text based on the sentiment of its nearest neighbors, which are determined based on a similarity metric such as Euclidean distance. K-NN does not make any assumptions about the underlying distribution of the data and can be effective in SA when the data is high dimensional and complex [1].

- **Decision Tree (DT)**

DT is a popular machine-learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences in a tree-like structure. In a DT, the dataset is split based on different attributes/features at each node of the tree. The algorithm selects the best attribute to split the data based on certain criteria (e.g., information gain or Gini impurity) to maximize the homogeneity or purity of the resulting subsets. The tree continues to grow by recursively splitting the data until a stopping condition is met. This can be a predefined depth limit, a minimum number of instances per leaf, or when there are no more attributes to split. Each leaf node represents a class or a predicted value for regression tasks. During the prediction phase, a new instance is traversed down the tree by following the decision paths based on the attribute values. The final prediction is determined by the class or value associated with the leaf node reached. DT have several advantages. They are easy to understand and interpret, as the resulting tree structure can be visualized. DT can handle both categorical and numerical data and are robust against noise and missing values. They can also capture non-linear relationships between features. However, DT are prone to overfitting, where they memorize the training data excessively. To overcome this, techniques like pruning or using ensemble methods like RF can be employed [3] (Al-Ghobesi, 2025).

- **Random Forest (RF)**

Random Forest (RF) is a popular ML technique that is used for both classification and regression tasks. It is an ensemble learning method that combines multiple DT to make predictions. In a RF, a collection of DT is created, where each tree is trained on a different subset of the data. To build each DT, a random subset of features is selected for each split, hence the term "random" in Random Forest. This random feature selection helps to introduce diversity among the trees and reduce the risk of overfitting. During the prediction phase, each DT in the RF independently makes predictions, and the final prediction is determined by majority voting

(in classification) or averaging (in regression) of the individual tree predictions. RF have several advantages. They can handle high-dimensional datasets with a large number of features, and they are robust against overfitting. They are also capable of capturing complex relationships between variables and handling missing data. Additionally, RF provide estimates of feature importance, allowing insights into the relative importance of different features in the prediction process (Omer, 2024, & Alasmari, 2023, Mleiki, 2025).

IV. Experiments and Results

In this study, we conducted SA of Arabic text using three ML techniques: K-NN, DT, and RF. We evaluated the performance of these techniques accuracy as the metric, and applied them to three different methods of labeling the data for ASA. These methods included manual labeling by humans, labeling by ChatGPT, and labeling by Bard Google. The goal of the experiment was to compare the performance of these techniques and methods and determine the most effective approach for SA of Arabic text. The accuracy scale was used to measure the performance of each method and technique, with higher accuracy indicating better performance. In the following section, 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 SA of Arabic text. The baseline models used in our experiments are in Section B. All the scripts constructed for the experiment are in RapidMiner. RapidMiner was used to create baseline models.

A. Evaluation criterion

To evaluate the effectiveness of a proposed approach, which involves using manual labeling by humans, labeling by ChatGPT, and labeling by Bard Google, we utilized one evaluation metrics in this paper. Four evaluation metrics have been utilized to evaluate ML models in this study. They are accuracy, precision, recall, and f-score measures. Before evaluating the effectiveness of the prediction model, the dataset was partitioned into testing and training sets. As shown in Eq (1), (2), (3), and (4).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+TN}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F} - \text{score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall+Precision}} \quad (4)$$

B. Baseline models

In the present study, we implemented four ML models as a baseline in our experiments. They are commonly used in the classification of approaches and constructed based on the training data. Examples include the K-NN, DT, and RF techniques.

C. Experimental results

Table 4 presents the results of different ML algorithms for SA, with each algorithm being labeled using different methods: manual labeling by humans, labeling by ChatGPT, and labeling by Bard Google، For the K-NN algorithm. Manual labeling by humans achieved an accuracy of 73.82%, while ChatGPT labeling achieved a slightly higher accuracy of 74.91%. Bard Google labeling had the lowest accuracy at 66.55%. In terms of recall, manual labeling had the highest value at 58.51%, followed by ChatGPT labeling at 56.82%. Bard Google labeling had a recall of 54.66%. For precision, both manual labeling and ChatGPT labeling had similar high values of 87.46% and 87.47%, respectively. Bard Google labeling had a precision of 81.15%. The F-score, which combines precision and recall, was the highest for manual labeling at 70.11%, followed by ChatGPT labeling at 68.88%. Bard Google labeling had an F-score of 65.32%.

TABLE 4. THE RESULTS ASA THROUGH MANUAL LABELING BY HUMANS, LABELING BY CHATGPT AND LABELING BY BARD GOOGLE USING K-NN TECHNIQUE.

Approaches	Accuracy	Recall	Precision	F-score
Labeling using manually by humans	73.82%	58.51%	87.46%	70.11%
Labeling By ChatGPT	**74.91%**	56.82%	87.47%	68.88%
Labeling By Bard Google	66.55%	54.66%	81.15%	65.32%

Table 5 presents the results of different ML algorithms for SA, with each algorithm being labeled using different methods: manual labeling by humans, labeling by ChatGPT, and labeling by Bard Google، For the DT algorithm. Manual labeling by humans achieved an accuracy of 57.45%, while ChatGPT labeling resulted in a slightly higher accuracy of 58.18%. Bard Google labeling had the lowest accuracy, with 53.45%. In terms of recall, manual labeling had the highest value at 44.93%, followed by Bard Google labeling at 41.18%. ChatGPT labeling had the lowest recall at 40.00%. For precision, ChatGPT labeling achieved the highest value of 52.52%, followed by Bard Google labeling at 50.44%. Manual labeling had a precision of 52.06%. The F-score, which combines precision and recall, was highest for manual labeling at 48.23%. Bard Google labeling and ChatGPT labeling had F-scores of 45.34% and 45.41%, respectively.

TABLE 5. THE RESULTS ASA THROUGH MANUAL LABELING BY HUMANS, LABELING BY CHATGPT AND LABELING BY BARD GOOGLE USING DT TECHNIQUE.

Approaches	Accuracy	Recall	Precision	F-score
Labeling using manually by humans	57.45%	44.93%	52.06%	48.23%
Labeling By ChatGPT	**58.18%**	40.00%	52.52%	45.41%
Labeling By Bard Google	53.45%	41.18%	50.44%	45.34%

Table 6 presents the results of different ML algorithms for SA, with each algorithm being labeled using different methods: manual labeling by humans, labeling by ChatGPT, and labeling by Bard Google. For the DT algorithm. Manual labeling by humans achieved an accuracy of 55.64%, while ChatGPT labeling resulted in a slightly higher accuracy of 57.09%. Bard Google labeling had the lowest accuracy, with 51.64%. In terms of recall, Bard Google labeling had the highest value at 35.96%, followed by manual labeling at 34.31%. ChatGPT labeling had the lowest recall at 33.67%. For precision, ChatGPT labeling achieved the highest value of 52.31%, followed by Bard Google labeling at 50.12%. Manual labeling had a precision of 51.72%. The F-score, which combines precision and recall, was highest for Bard Google labeling at 41.87%. Manual labeling and ChatGPT labeling had F-scores of 41.25% and 40.96%, respectively.

TABLE 6. THE RESULTS ASA THROUGH MANUAL LABELING BY HUMANS, LABELING BY CHATGPT AND LABELING BY BARD GOOGLE USING RF TECHNIQUE.

Approaches	Accuracy	Recall	Precision	F-score
Labeling using manually by humans	55.64%	34.31%	51.72%	41.25%
Labeling By ChatGPT	**57.09%**	33.67%	52.31%	40.96%
Labeling By Bard Google	51.64%	35.96%	50.12%	41.87%

.Fig. 2. This graph compares the accuracy of different labeling methods across various algorithms.

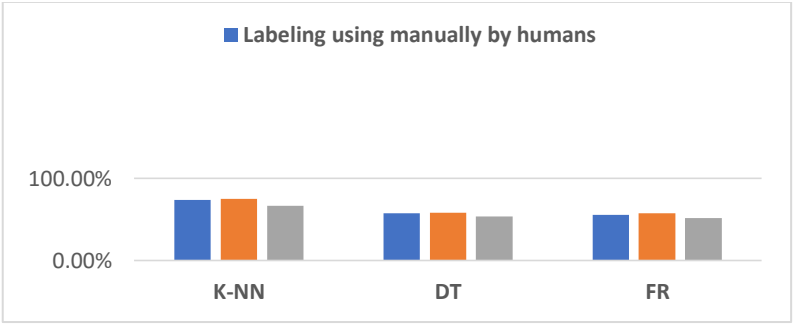

Fig. 2. A graphical representation of the Accuracy metrics for manual labeling by humans, labeling by ChatGPT and labeling by Bard Google.

V. conclusion

This study is dedicated to analyze the Arabic sentiments of manually collected Arabic dataset. It related to mobile applications for Yemeni universities. The SA tasks were based on the reviews collected from Google Play Store. The machine learning models used were K-NN, DT, and RF. Thus, the advantages and disadvantages of a particular mobile application can be seen based on its users' reviews. Several mobile applications in Yemeni universities providers can concentrate on correcting defects and better adjustment to the demands of their users. These methods included manual labeling by humans, labeling by ChatGPT and labeling by Bard Google. Based on the experimental findings, the K-NN technique demonstrated superior performance in Arabic sentiment analysis (ASA) using ChatGPT models, achieving an accuracy of 74.91%. Furthermore, the utilization of the proposed active labeling method with ChatGPT resulted in higher accuracy compared to other labeling methods. The study proposes that combining the K-NN technique with ChatGPT models and employing the suggested active labeling method are effective approaches for ASA using ChatGPT. The empirical results highlight the promising outcomes of the machine learning-based approach in evaluating students' opinions regarding higher education institutions. For future work, expanding the quantity of data (big data) and incorporating Arabic and English reviews are recommended.

References

[1] Alasmari, J. S. . (2023). The Dynamics of Verbal and Non-Verbal Linguistic Communication in The Saudi Sports Community. *Arts for Linguistic & Literary Studies*, *5*(4), 539–569. https://doi.org/10.53286/arts.v5i4.1676

[2] Al-Ghobesi, A. A. H. (2025). Risks of Relying on Artificial Intelligence in Learning Arabic Language Sciences Through the Meta Application. *Arts for Linguistic & Literary Studies*, *7*(1), 396–419. https://doi.org/10.53286/arts.v7i1.2420

[3] Al-Hagree, S., & Al-Gaphari, G. (2022). "Arabic Sentiment Analysis Based Machine Learning for Measuring User Satisfaction with Banking Services' Mobile Applications: Comparative Study". In 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA) (pp. 1-4). IEE .

[4] Al-Hagree, S., & Al-Gaphari, G. (2022). 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) (pp. 1-6). IEEE.

[5] Al-Shalabi, A. A., Al-Gaphari, G. (2023). Salah, A. H., & Alqasemi, F. Investigating the Impact of Utilizing the K-Nearest Neighbor and Levenshtein Distance Algorithms for Arabic Sentiment Analysis on Mobile Applications. JAST, 1(2).

[6] Baragash, R., & Aldowah, H. (2021). Sentiment analysis in higher education: a systematic mapping review. In Journal of Physics: Conference Series (Vol. 1860, No. 1, p. 012002). IOP Publishing.

[7] Mleiki, A. K. (2025). Exploring Saudi EFL Learners' Perspectives on Digital Writing Tools for Mitigating Emotional Challenges in Foreign Language Writing. *Arts for Linguistic & Literary Studies*, *7*(1), 577–590. https://doi.org/10.53286/arts.v7i1.2373

[8] Motlagh, N. Y. (2023). Khajavi, M., Sharifi, A., & Ahmadi, M. The Impact of Artificial Intelligence on the Evolution of Digital Education: A Comparative Study of OpenAI Text Generation Tools including ChatGPT, Bing Chat, Bard, and Ernie. arXiv preprint arXiv:2309.02029.

[9] Omer, N. I. M. (2024). Maintaining Meaningful Human Interaction in AI-Enhanced Language Learning Environments: A Systematic Review. *Arts for Linguistic & Literary Studies*, *6*(3), 533–552. https://doi.org/10.53286/arts.v6i3.2083

[10] Praveen, S. V., & Vajrobol, V. (2023). Understanding the perceptions of healthcare researchers regarding ChatGPT: a study based on bidirectional encoder representation from transformers (BERT) sentiment analysis and topic modeling. Annals of Biomedical Engineering, 1-3.

[11] Susnjak, T. (2023). Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. arXiv preprint arXiv:2302.06474.

[12] Toçoğlu, M. A., & Onan, A. (2021). Sentiment analysis on students' evaluation of higher educational institutions. In Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020 (pp. 1693-1700). Springer International Publishing.

[13] Wang, Z., Xie, Q. (2023). Ding, Z., Feng, Y., & Xia, R. Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339.

[14] Zhu, Y., Zhang, P., Haq, E. U., Hui, P., & Tyson, G. (2023). Can chatgpt reproduce human-generated labels? a study of social computing tasks. arXiv preprint arXiv:2304.10145.