From Accuracy to Insight: Explainability in Review Rating Prediction with Transformers

Dhefaf T. Radain, Dimah Alahmadi, Arwa M. Wali
Department of Information Systems-Faculty of Computing and Information Technology,
King Abdulaziz University, Saudi Arabia

Abstract—**Mobile application (app) reviews provide valuable information that facilitates understanding of users' needs, leading to better design of developed products. They have abundant data that can be utilized by different models to explain the prediction results to stakeholders. This will lead mobile app developers to trust and rely on the models that are used to develop their apps and satisfy the users' needs. To leverage this information, outstanding improvements in complex learning algorithms have led to the development of transformer-based models that are used for natural language processing (NLP) and to exploit rating predictions. However, such models are complex and lack explainability, especially for Arabic reviews. Most studies have applied explainability models for transformer-based models to the English language and various other languages but not the Arabic language. This study presents a rating prediction explainability (RPE) framework that combines transformer-based and explainability models for review rating predictions from mobile government (m-government) apps. The transformer-based models predict the ratings for reviews written in English or Arabic. Then, local explainability models, such as SHapley Additive exPlanation (SHAP) and local interpretable model-agnostic explanations (LIME), explain and visualize the results. In RPE, not only high prediction accuracy was achieved for both English and Arabic reviews, but the resulted predictions were also justified with consistency between the different explainability models. The transformer-based model ELECTRA yielded the highest accuracy and F1 score of 96% for the rating prediction of English reviews, whereas the transformer-based model AraBERTv2 had 95% accuracy and F1 score for the rating prediction of Arabic reviews. The results of both explainability models provided equivalent explanations and emphasized the same words that affected the predicted ratings.**

Keywords—Explainability; LIME; review rating prediction; SHAP; transformer-based models

I. Introduction

In recent years, an increasing number of user reviews have been posted on various internet services. These services include social media, e-commerce, and microblogs [1]. User reviews are comments and opinions provided by consumers about a product, business, or application (app) [2]. They consist of essential information that can capture the users' preferences over multiple items [3]. They provide pivotal information regarding the quality of the item being reviewed to help others decide whether to buy the product, deal with the business, or download the app [4]. This has prompted recent studies in this field on complementing item representation and improving recommendation performance [3].

One source of these reviews is mobile apps. They are a convenient way for users to share their opinions and thoughts. The interaction between users and mobile apps under several usage contexts have made the need for better app design a crucial issue. This presents a considerable challenge for app developers and marketers in developing suitable apps for users on the basis of their needs. Constructive and timely user reviews help developers become aware of user needs, creating opportunities for innovation [5]. The star ratings and the users' written reviews that is provided by the apps create an understanding of the consumers' needs and preferences, which leads to enhancements in the apps' features or the development of more advanced apps [6].Furthermore, one of the most important types of reviews that require careful analysis is related to mobile government (m-government) apps. The evolution of information and communication technology (ICT) has led governments worldwide to digitize their services using mobile technology, creating the concept of m-government. M-government allows for better communication with citizens/residents and provides them with better access to services [7]. The number of mobile phone users worldwide has reached 6.378 billion [8], and in Saudi Arabia, the number of smartphone users is expected to increase to 36.17 million in 2025 [7]. An evaluation of the current usage of m-government apps will help determine their effectiveness [6].

Turning this unstructured feedback into actionable insights requires natural-language models that are both accurate and linguistically inclusive. Numerous approaches have leveraged user reviews to extract knowledge related to their content or semantics in several languages [9]. To extract and work within user reviews, studies have utilized the features provided by deep learning (DL) in natural language processing (NLP). Among DL architectures, transformer-based models are considered important landscapes for NLP. On many tasks, they outperform many other state-of-the-art DL models [10] [9]. This is because, with traditional models, training incurs a high computational cost, which limits the volume of tuning that could be done on a model [11]. Furthermore, the size of the dataset used to train a model limits the ability to measure how the model is advancing. However, since transformer-based models are already trained on enormous datasets, this reduces the computations needed for the training process. In addition, they eliminate the need for an enormous dataset to obtain significant results [12].

On the other hand, the black-box nature of these models is considered a shortcoming. This means that the internal operations of the classifiers are unknown to humans [13]. Understanding the reasons behind the predictions made by a model is quite important for taking action based on these reasons [14]. Explainability is the ability of humans to understand a model and its behavior. In some cases, explainability is considered a prerequisite for accepting the results of the model [15]; for example, the use of the model in medical applications or terrorist recognition cannot be based on blind faith, considering the consequences. Furthermore, any predictions related to a text should be linked with an explanation that indicates how and why the predictions were made. This type of understanding can be achieved by tracing each decision made by a model back to individual words [16].

Studies have shown that explainability models, such as SHapley Additive exPlanation (SHAP) and local interpretable model-agnostic explanations (LIME), are useful for adding explanations when incorporated into model workflows [17]. Recently, various studies have used SHAP as an explainability model. For example, it was used to measure the contributions of specific tokens toward specific predictions of material properties [18]. In addition, it was used as an explanation method for text classification developed by the BERT transformer-based model [19]. LIME was used as a basic explainability model for detecting patients with specific diseases on the basis of the predictions of several transformer-based models [20]. Additionally, it was used to explain the reasons behind the identification of harmful content in tweets [21].

The intersection of highly rated apps and public services makes m-government apps an ideal, high-impact test bed for such explainable techniques. The use of m-government in Saudi Arabia began to take shape in different sectors of the country [22]. Predicting the reviews' ratings and providing an explanation for the predicted ratings will enrich the Saudi Arabian app development market to develop apps based on users' demands. However, these reviews frequently mix modern standard Arabic with colloquial dialects which are underrepresented in standard corpora. Arabic language presents unique challenges that are not present in English-based NLP pipelines. This is due to the language's rich morphology, complex grammar, data scarcity, and variety of dialects [23], that result in inconsistent predictions and inaccurate and unreliable models. Even with the growing number of advancements and improvements in transformers and explainability models, there is still a lack of studies that have used such models in Arabic NLP tasks, especially in the field of predicting review ratings.

Consequently, this study aims to provide the Rating-Prediction-Explainability (RPE) framework, which combines various transformer models with SHAP and LIME to deliver both accuracy and transparency in predicting the ratings of m-government apps in Saudi Arabia based on written reviews in English and Arabic.

The main objectives of this study are as follows:

- Creating a dataset for m-government apps for both English and Arabic language.
- Developing a review rating prediction framework for English and Arabic reviews from m-government apps using Transformer-based models.
- Providing explainability for the resulting prediction from the transformer-based models using explainability models.
- Proving the consistency in the explainability between the different explainability models used in the framework.

The remainder of this study will discuss the related works to this study from the literature, as well as highlight the research gap. Next, it discusses the details of the methodology used for the proposed model, the data, and the experiment and evaluation of the model. Then it presents the results and discusses them. Finally, the study ends with a conclusion and provides limitations and future work.

II. Related Work

Given the importance of user-generated reviews, developers and academic researchers have conducted several studies to determine the best method for predicting ratings based on the review content and the language of the review. This has resulted in a high demand for understanding the predicted results [24]. This section highlights the different studies that have focused on rating prediction on the basis of users' reviews and their language, transformer-based models, and different explainability models

A. Rating Prediction

Because reviews and their associated ratings are valuable and not always available, rating prediction has been developed to generate ratings for non-rated items [25]. However, discovering and extracting the users' needs is a nontrivial task because only one-third of app reviews contain objective statements. Moreover, processing many unstructured reviews to extract possible user needs can be tedious. Therefore, automatic extraction is more efficient [5].

Consequently, developers and academic researchers are working on improving the performance of prediction models by extracting diverse features, such as lexical patterns, words, semantic topics, and syntactic structure, from the contents of reviews [26].

Current technical innovations in the field of machine learning (ML) have led to its utilization in various scenarios, including rating prediction [27]. Moreover, DL-based models are effective not only in improving performance but also in learning feature representations from scratch [4], such as Qiao et al. [5] who proposed a domain-oriented approach based on DL that combines a convolutional neural network (CNN) and a recurrent neural network (RNN). The model classifies user feedback to identify the type of information within the user reviews of mobile apps. The results of the study revealed that the proposed model yielded more valuable information, such as essential key-words and more consistent topics. In addition, this DL approach outperformed the traditional classification methods because it captures more contextual and semantic relationships between words.

Moreover, Ahmed and Ghabayen [28] proposed a DL bidirectional gated recurrent unit (Bi-GRU) architecture model for rating prediction. It consists of two phases: the first phase involves polarity prediction, whereas the second phase involves the prediction of review ratings from the text of a review. The results of their experiment indicate that the proposed framework can significantly improve rating prediction compared with the baseline methods. Another study that used a CNN to predict ratings for mobile apps was [29] by Aslam et al. The classifier extracts textual and non-textual information from each app review. Textual information is preprocessed for each review, and a digital path is created. Finally, a prediction is generated for the app review. The results of the proposed approach revealed a significant improvement in rating prediction for reviews.

However, the rating is not entirely determined by the content of the review because the user may give a low rating but write a positive review. Consequently, different approaches and models have been proposed for rating prediction [26]. Sadiq et al. [30] proposed a DL-based framework for predicting the contradictions between rating predictions in Google Apps. The polarity of reviews is predicted using sentiment analysis to determine the ground truth. Next, DL models are trained on the ground truth to predict star ratings from text reviews. The results of their experiment showed that unbiased star ratings could be predicted based on actual reviews written by users [30]. Table I provides a summary of the proposed models along with their strengths and limitations.

B. Transformer-Based Models

The effectiveness of DL models when a large amount of data is used has been demonstrated; however, there are many limitations related to them. This caused the researchers to focus increasingly on much more sophisticated models that capture the dynamics of the used language. For example, RNN can effectively capture the contextual information and long-term dependencies in the text, however, there are several challenges related to long sequences which makes it difficult to figure out the correlations between the words. Moreover, CNN is known to be used for text classification by extracting local features from the provided input. This allow easy detection of text patterns. However, it provided low accuracy in some of the cases [31]. Additionally, there are some cases in which the available data are not sufficient to train a classifier to obtain justifiable results [32], and training incurs a high computational cost, which limits the model training process [12]. Consequently, transformer-based models were introduced with pretrained models that provide a vast amount of data and reduce the computational cost needed for training DL models [15].

Given the vast amount of information available in users' reviews available online, Kaur and Kaur [33] classified app reviews as relevant, feature requests, or bug reports. They used BERT to extract the contextual connections between written reviews. To assess the efficiency of their proposed model, five datasets were tested, and the model outperformed state-of-the-art models.

Additionally, Shiju and He [34] built a classification model via BERT, XLNet, RoBERTa, ELECTRA, BioBERT, and ALBERT transformer-based models. The models classified drug ratings based on the written textual reviews. This study identified reviews that are inconsistent with the given ratings. Moreover, BioBERT outperformed the other models because it specializes in medical data, which was the focus of this study.

Moreover, Chowdhury et al. [35] investigated the effect of BERT on Arabic text classification by incorporating formal and informal text in the training process. The goal was to classify text into categories such as 'sports', 'human rights', and 'politics', among others using BERT. The study revealed that training with formal text was more generalizable than training with informal text while enabling the correct classification of text. The formal text dataset consisted of social media posts from popular Arabic news channels that cover YouTube, Facebook, and Twitter, whereas the informal text was from Arabic tweets from popular accounts. The study revealed an overall improvement in text classification effectiveness.

In terms of rating prediction, Liu [9] predicted restaurant ratings based on textual reviews from the Yelp dataset via ML models such as Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Linear Support Vector Machine (SVM) and transformer-based models such as RoBERTa, DistilBERT, BERT, and XLNet. Finally, the authors compared the different models in terms of computational resources, speed, and evaluation metrics. The study showed that transformer-based models had higher accuracy than ML models did. One of the reasons for this result is that transformer-based models are pretrained on very diverse data. This diversity improves the generalizability and robustness of these models [36]. Table II provides a summary of the transfromer-based models along with their strengths and limitations.

C. Explainability Models

Recently, breakthroughs in the use of DL models have led to improvements in their accuracy and utilization in several scenarios [27]. However, these models are still considered black-box models [16], and their growing complexity makes reliable, fair, and safe models mandatory [37]. To address this concern, explainability models have been developed. They can be applied to any classification task [19]. In prediction-based models, understanding decisions is relevant for evaluating the reliability of the resulting predictions and detecting possible biases in the models [38].

Owing to the increasing complexity of DL models, reliable, fair, and safe models are needed [37]. Explainability models aim to increase trust by ensuring the transparency, dependability, and stability of the model used [13].

Research have shown that SHAP and LIME explainability models can be added to the workflow of a model used to explain the resulting predictions [17]. As a result, these models have been applied in different domains to enhance explainability. Ahmed et al. [39] developed a model that integrates LIME and BiLSTM to detect fake news. In this model, BiLSTM serves as the classification model for fake news detection, and LIME is used as the explainability model. The model achieved high accuracy and outperformed other ML models.

In the medical domain, Ilias and Askounis [20] used LIME to explain the ability of a BERT transformer-based model to predict whether a patient is suffering from Alzheimer's disease (AD) on the basis of the patients' words used while speaking in English. This study re-vealed significant language differences between patients with and without AD. These differences were explained visually, revealing the diverse patterns of words, expres-sions, and verb tenses used by patients.

Similarly, Rao et al. [40] predicted the incidence of heart failure using the BEHRT transformer-based model. This model was applied to data from patients' electronic health records in the UK. The authors provided explanations for the resulting predictions. BEHRT worked robustly with large-scale, sequential data and outperformed other traditional DL models. The

TABLE I. SUMMARY OF RATING PREDICTION RELATED WORK

Ref	Model	Strengths	Limitations
[5]	CNN and RNN	Capturing more semantic and contextual relationship between the words	The use of one public data source and one method of analysis
[28]	Bi-GRU for rating prediction based on DL	Evaluate the framework on real-world datasets	The large number of unique words causes inefficient encoding for them
[29]	App review classification approach based on CNN	Works on textual and non-textual information	Considering a limited number of reviews
[30]	Neural network–based method for understanding and predicting users' rating behaviors	Unify aspect rating with the review content	Detecting fake reviews

TABLE II. SUMMARY OF TRANSFORMER-BASED MODELS RELATED WORK

Ref	Model	Text Type	Strengths	Limitations
[33]	BERT	App reviews	The model achieved high performance in prediction for both long and short datasets	The used dataset had some internal validity issues
[34]	BERT, XLNet, RoBERTa, ELECTRA, BioBERT, and ALBERT	Drugs reviews	Identifying the inconsistent reviews with given ratings	Relying on the drug name for the classification.
[35]	BERT	Social media posts	The classification into multiple labels	Collecting and pre-training a large set of data is impractical
[9]	ML models (NB, LR, RF, and Linear SVM) and transforrmer-based models (RoBERTa, DistilBERT, BERT, and XL-Net)	Restaurant reviews	The experiment with several transformer-based models	No much difference in the accuracy between the ML models and the transformer-based models.

authors used the perturbation concept as a local surrogate method to explain the resulting predictions. They quantified the contribution of selected patient encounters and tested how the combination of the diagnosis and the medications affected the model's predictions.

Although transformer-based models have achieved very promising results in several areas, understanding their results is still an area to explore [39]. Korangi et al. [41] proposed an innovative model representation leveraging a BERT transformer-based model for predicting credit risk for mid-cap companies. They referred to this approach as the transformer encoder for panel-data classification (TEP). The authors incorporated differential training with a multichannel architecture, leading to superior performance and improvement over traditional models by efficiently training with all the data. SHAP was applied to express the importance of ranking the different sources of data and their relationships.

As stated in the previous literature, the explainability of a model is an important factor for its reliability. Different types of text were used under different classification and interpretation models, as shown in Table III, which provides a summary of the explainability models.

III. RESEARCH GAP

To the best of our knowledge, even with many transformer-based models that have been used for review rating prediction and a respectable number of scientific studies that have focused on explainability in different fields, there is a lack of studies that combine these models with explainability to justify the predicted ratings. There is still a need to conduct more research in the field of review rating prediction, especially for mobile apps. They have abundant data that can be utilized by different models to explain the prediction results to stakeholders. This will lead mobile app developers to trust and rely on the models used to develop their apps and satisfy the users' needs. In addition, most studies have applied explainability models for transformer-based models to the English language and various other languages but not the Arabic language. Consequently, this study applies transformer-based models for review rating prediction in both the English and Arabic languages and then applies explainability models to justify the predicted ratings to help the app developers recognize the users' needs for the apps and understand the difficulties they face.

IV. RPE FRAMEWORK

In this study, the RPE framework is proposed to explain the predictions of the review rating of mobile applications generated using transformer-based models. The model consists of three modules: 1) a dataset creation module, where the dataset is collected, cleaned, split, and balanced; 2) a rating prediction module, where the input sequences are processed through three different layers to predict ratings via transformer-based models; and 3) an explainability module, where the resulting predictions are justified via visualization techniques of explainability models. The high-level architecture of the model is shown in Fig. 1.

A. Dataset Creation

The initial dataset consisted of user reviews collected from m-government apps. Reviews were collected from a set of apps based on their purpose. The dataset was then cleaned and split based on the detected language. Dataset Description Section discusses the details of this process. The reviews from the resulted datasets are then passed through the prediction layer to perform the rating prediction.

B. Prediction

The prediction module consists of three different layers: an embedding layer, a transformer layer, and a prediction layer. The rating is predicted for an app review by passing the input

TABLE III. SUMMARY OF EXPLAINABILITY MODELS RELATED WORK

Ref	Classification Technique	Explainability Model	Aim	Type of Text
[39]	BiLSTM	LIME	Analyzing and comparing the computational overhead	Covid-19 news
[20]	TEP	SHAP	The use of two explainability models to confirm their results	Market or pricing data
[40]	BERT, ALBERT, BioBERT, RoBERTa, XLNet, ConvBERT, and BioClinicalBERT	LIME	Detecting AD in patients based on the used words when speaking English	English spoken words
[41]	MatBERT	LIME and SHAP	predictions of material properties	Material properties written in English

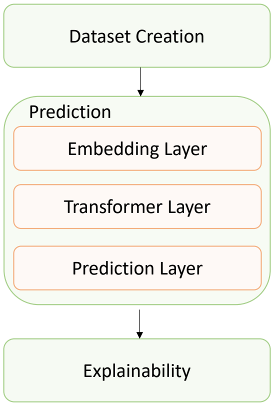


Fig. 1. RPE high-level architecture.

through these layers, which convert the app review into a format that is understandable by the machine and prepares it for the training process. In the following subsections, we provide detailed explanations of these layers.

1) *Embedding layer:* The transformer-based model receives the user review as discrete input tokens (sentence). The review cannot be processed directly as raw text; therefore, it passes through an embedding layer. The embedding layer converts the input into continuous, high-dimensional vector representations where each token is assigned a unique vector. Next, the model generates contextual embeddings by making some adjustments to the vector where tokens with similar meaning are brought together, allowing the model to capture the semantic relationship between them. As a result, embeddings are generated for each word within different contexts. The self-attention mechanism in the transformer-based models starts with the embedding vector. Then it refines the initial representation based on the context of the entire input [42], [43].

This layer considers text through three different embeddings: token embedding, sentence embedding, and positional embedding. In token embedding, the input is tokenized using the WordPiece tokenization method. The text input is converted into a numerical representation that is understandable by the machine (e_x1, e_x2, e_x3..., e_xL). In addition, additional tokens are added to each input sequence, such as <CLS>, which indicates the beginning of the input, and <SEP>, which separates sentences. The resulting tokens are then passed through the sentence embedding, where the words of each sentence are assigned the same embedding. This is accomplished by adding a learned embedding to each word to indicate the sentence to which the word belongs. Ultimately, an input sequence consisting of multiple sentences is packed together as a single sentence and passes through positional embedding. Each word in positional embedding is assigned a number that represents its position within the entire input (e_1, e_2, e_3, . . . , e_L). If a word is repeated within the input, the same number is assigned. Finally, the sum of the three embeddings generates a representation of a single shape that is passed as an input for the transformer layer. These embeddings contain the meaning and characteristics of the input word, which enables the model to understand their relation to the corresponding rating to perform the rating prediction on the testing dataset [42], [43]. A detailed view of this layer is shown in Fig. 2. We refer the readers to [42], [43] for a comprehensive explanation.

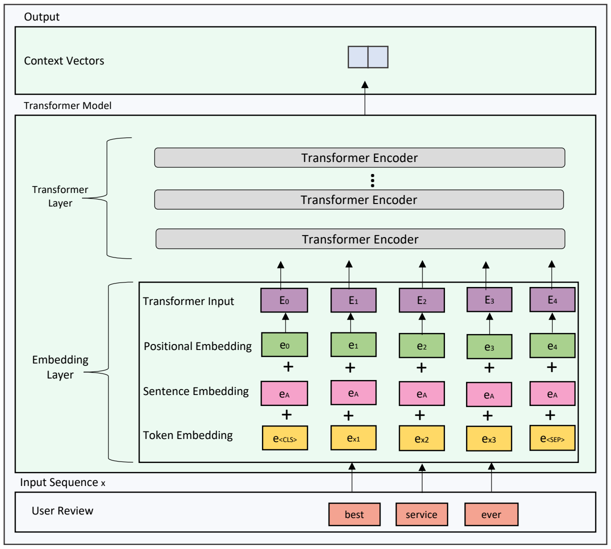


Fig. 2. Embedding and transformer layers (inspired by [43]) showing the flow of the user review as the input sequence into the output context vector.

2) *Transformer layer:* In the transformer layer, the input passes through the attention mechanism, normalization, and feed-forward layers, where the actual classification of the text occurs to extract the key characteristics [44]. First, the model takes the ¡CLS¿ token, which is considered the classification token, as the first input, followed by the remaining words in the input sequence. The input is then passed through the above layers. Each layer applies self-attention and passes the results through a feedforward network that passes the input again to the next layer above [11]. The final output of this layer, which represents the context vector of the input sequence, is then passed to the prediction layer.

3) Prediction layer: The output vectors of the transformer layer represent the important features provided by the input sequence. These features are then passed through a dense neural network to produce the logits. The logits are the output of the last neural network layer in raw unnormalized form before they are converted into probabilities. These raw scores are essential, as they encapsulate the model's initial predictions. Logits are then converted into probabilities by passing them through the Softmax functions in the Softmax layer. This transformation is crucial for making the model's output actionable and explainable. The softmax function maps the logits from the range of (-infty) to (+infty) to a range of 0 to 1, to ensure that the output values sum up to 1 and can be interpreted as probabilities. Lastly, these probabilities produce the predicted rating, as shown in Fig. 3 and passed through to the explainability layer along with the transformer-based model used for rating prediction.

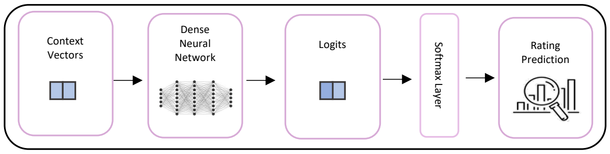


Fig. 3. Workflow of rating prediction layer.

C. Explainability

In the RPE framework, the SHAP and LIME explainability models are used to explain the predicted ratings.

1) SHAP: SHAP is a model-agnostic method used to explain the behavior of a model across all instances [45]. It identifies and prioritizes features that define classifications and predictions of any model by assigning a SHAP value to each feature [38]. In this work, the goal of using the Shapley value is to distribute the contributions between the features because they result in a certain prediction. The Shapley value of a feature (represented by a token) is the contribution of the feature to the model prediction [17].

In the RPE framework, SHAP receives the user review and transformer-based model as input and then creates N new samples from the same review via perturbation. The perturbed samples are created by masking random words from the original user review. These samples are fed into the model to generate their prediction. The difference between the predicted rating of each perturbed sample and the predicted rating of the original sample represents the contribution of the masked word to the prediction. The SHAP value of a feature reflects the strength of the impact of the feature on the resulting prediction via the model. Thus, it can be used as an importance score for features [17].

2) LIME: LIME is a model-agnostic explanation method that generates a local explanation for a complex model by justifying its decision for a specific observation. This approach implements a linear explainability model trained to estimate the prediction of a black-box model [46]. It works via a direct approach of training the model locally on specially generated representations that are different from the original input sequence of the model and observing the changes that occur in the prediction [47].

In RPE, LIME creates N new perturbed samples from the input review, similar to SHAP. These samples are fed to the transformer-based model to generate the predictions. Moreover, LIME weights these perturbed samples based on their closeness to the original sample to indicate their relative importance. Then, a linear regression model is trained on the perturbed samples and the weighted samples to explain a specific prediction. The coefficients are extracted from the linear model and used to explain the local behavior of the model. If the coefficient is positive for a specific word, this word positively affects the model's prediction, whereas a negative coefficient indicates the opposite. The degree of the effect on the prediction depends on the value of the coefficient [16].

V. Dataset Description

In this section, all the processes that the data underwent are described, from collection to cleaning, balancing, and splitting to create the final datasets. The dataset creation process is illustrated in Fig. 4.

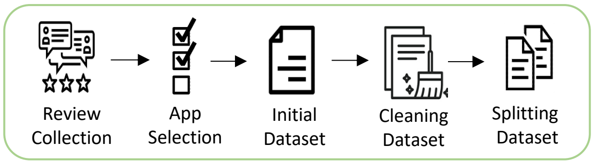


Fig. 4. Initial dataset creation.

A. Data Collection

The data needed for this study was users' reviews and ratings from m-government mobile apps. And since the government of Saudi Arabia is moving towards providing citizens and residents with additional access to governmental information and services through the m-government apps [22], its certified m-government apps platform was used to collect the data and create the dataset for the study. This certified platform is called the Saudi Unified National Platform. It includes the services related to healthcare, Hajj and Umrah, education and training, tourism,culture, and entertainment, safety and environment, Islamic affairs, business and entrepreneurship, and others. In the end, the dataset consisted of real-world data of users' reviews and ratings of m-government apps that provide social services for citizens and residents in Saudi Arabia.

B. Data Cleaning

Since rating prediction in this study is done using transformer-based models, the usual text preprocessing is not needed, such as removing null values, stop words, white spaces, punctuation marks, ASCII, duplicate letters, contractions, hashtags, emojis, and URLs. This comes from the fact that the transfer learning techniques are feature independent, which means that their power to understand natural language comes from their ability to contextually model the language. When they are applied to textual data that is heavily cleaned, they become worse because of the loss of contextual information needed by the model [48]. The dataset was cleaned by removing unnecessary attributes, that is, the time of the review, name of the user, title of the review, and version of the app, from each review which were collected by the data collection tool. Afterward, reviews were examined manually and all repeated reviews by the same user were deleted, as well as were reviews that were not in Arabic or English. Moreover, reviews that did not provide useful information for the study were eliminated. Furthermore, the reviews that were written in multiple languages were revised, and only one language for the review was retained either English or Arabic. In addition, words written in a language different from the review are translated or deleted if they are not necessary. Table IV lists examples of the cleaning processes.

TABLE IV. Testing Reviews Before and After Cleaning

Review Before Cleaning	Review After Cleaning	Problem
كتبت البلاغ لاكثر من مرة ولم يصلني رقم التفعيل sms	كتبت البلاغ لاكثر من مرة ولم يصلني رقم التفعيل رسالة نصية	Mixed language
وكلتكم باخراج زكاة فطري لي ولاهل بيتي	–	non-useful information
wonderful charity app. Hoping to add search engine التطبيق رائع ..و أتمنى إضافة آلية البحث	wonderful charity app. Hoping to add search engine	Multiple languages
नयाल	–	Non-English nor Arabic

C. Data Splitting

The RPE framework explains rating predictions for both English and Arabic reviews. Consequently, the initial dataset, which contains reviews in both languages, was split into two datasets based on the language of the reviews to produce the final datasets. The English and Arabic datasets consisted of 4381 and 6509 reviews, respectively, which were distributed among different class ratings from 1 to 5, as shown in Table V.

TABLE V. Actual Rating Distribution of M-Government Apps Reviews that Provide Social Services for Citizens and Residents in Saudi Arabia

Dataset	Rating Class				
	1	2	3	4	5
English	2581	320	254	224	1002
Arabic	4410	508	390	259	942

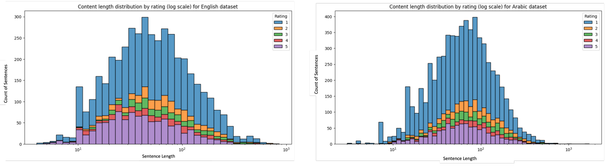
D. Data Balancing

In classification problems, the performance of the DL model becomes worse when dealing with imbalanced data because the model tends to learn more towards the majority class than the minority class. This degrades the performance of the model and reduces the ability to generalize the results because of the difficulty in accurately predicting the minority class in the presence of the majority classes. To address the imbalanced numbers of reviews among the rating classes, as shown in the previous table, the datasets were examined manually, and the main observations for minority classes 2, 3, and 4 were related to the quality of the written reviews:

- Some of the reviews in these rating classes were meaningless and included only emojis, unknown words, reviewer names, etc., as shown in Table VI. These reviews do not highlight the features in the review to distinguish between them and know which of them has a stronger impact on the given rating than others.
- Users tended to write long reviews when they were completely satisfied or completely disappointed with an app. Consequently, the review lengths of these minority rating classes were noticeably shorter than those of classes 1 and 5. Fig. 5 shows the difference in sentence length across the rating classes.
- Because users could not provide a star rating unless they also provided a written review, they sometimes wrote non-useful reviews only to reach the point of a star rating, as mentioned by some of the users.

TABLE VI. Data Sample of M-Government Apps Reviews that Provide Social Services for Citizens and Residents in Saudi Arabia

Source	Content	Rating
NWC	Thx	3
Sehhaty	Gd r go bjy zoo zlay e ej tzzzI eeu h	2
Tawakkalna	أبو هيثم	4
Bader	Saeed1234	1
Citizen Account	🇸🇦 ♡	5



a) Sentence Length Distribution for English Dataset

b) Sentence Length Distribution for Arabic Dataset

Fig. 5. Sentence length of m-government app reviews that provide social services for citizens and residents in Saudi Arabia.

All of the above may affect the model during the training process, as these factors may prevent understanding of various aspects of reviews in the minority rating classes, thus resulting in the failure to predict their ratings correctly. This led to the removal of the reviews that belonged to class 2, 3, and 4, and only the reviews from class 1 and 5 were used in the study to train the rating prediction model. Thus, this study focused on the polarity of reviews to measure the satisfaction of app users and to understand their needs for app development.

However, these two rating classes were also unbalanced. Therefore, a text augmentation technique was used to add data to the minority class. Additional data was added to the minority class from the unused collected data. Additional reviews were collected from other apps on the platform, such as health and education apps. Reviews of rating classes 1 and 5 were added to the English dataset to increase the amount of data and balance the two rating classes. For the Arabic dataset, only reviews of rating class 5 were added, and rating class 1 was down-sampled to balance the two classes. The final datasets consisted of 5430 reviews from 67 applications for the English dataset and 5470 reviews from 82 apps for the Arabic dataset.

Table VII shows the distribution of the reviews among the rating classes.

TABLE VII. FINAL DATASET

Dataset	Rating Class	
	1	**5**
English	2709	2723
Arabic	2791	2690

VI. EXPERIMENT AND EVALUATION

In this section, all the experimental settings used for the transformer-based models and the explainability models are listed. In addition, the evaluation metrics for both types of models are highlighted.

A. Experimental Settings

Details regarding the implementation of the transformer-based models used in this experiment, training strategy, hyperparameters, and model performance for each dataset are provided below:

1) Implementation details: Transformer-based models were implemented in the same manner, as indicated in [43]. To use the models in text classification, the problem was considered a rating prediction problem in which only the encoder part was used. The dataset was split into 80/10/10 training/testing/validation data.

2) Training strategy: This study used the same training strategy, as in [43]. The model iterated through different hyperparameters and transformer-based models to find the best performing model for each dataset in the classification task. The hyperparameters used for both datasets are listed in Table VIII.

3) Prediction models: Several transformer-based models, namely, BERT, XLNet, RoBERTa, ELECTRA, ALBERT, XLM_RoBERTa, MARBERT, QARiB, and AraBERT, were fine-tuned on both datasets in this experiment to obtain the highest accuracy and F1 score.

4) Explainability models: SHAP and LIME were applied to the testing datasets using the transformer-based models, which resulted in the highest accuracy and F1 score.

TABLE VIII. MODEL HYPERPARAMETERS

Training Argument	Value
Training Batch Size	16
Evaluation Batch Size	16
Number of Epochs	4
Evaluation Steps	300
Learning Rate	5e-5
Output Shape	(768, 2)

B. Evaluation Metrics

Two types of evaluations were conducted, namely, prediction model and explainability model evaluations, as follows:

1) Prediction model evaluation metrics: The classification metrics used to measure the overall performance of the transformer-based prediction models were accuracy and the F1 score.

- Accuracy: the ratio of the number of correctly predicted ratings to the total number of predicted ratings [49]. It is calculated by Eq. (1):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, TP is true positive which is the number of outcomes where the model correctly predicts the positive class, TN is true negative which is the number of outcomes where the model correctly predicts the negative class, FP is false positive which is the number of outcomes where the model incorrectly predicts the positive class, and FN is false negative which is the number of outcomes where the model incorrectly predicts the negative class.

- F_1 score: the average rate between the recall and precision [49]. It is calculated by Eq. (2):

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2}$$

where, precision is the proportion of correct positive identifications out of all positive identifications and recall is the proportion of correct positive identifications out of all true positive instances. They are calculated via Eq. (3) and Eq. (4):

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

2) Explainability model evaluation metrics: Despite the massive amount of knowledge established around explainability models, there is no general agreement in the literature on how they should be defined or evaluated [50]. The literature suggests different evaluation metrics, such as comprehensibility, accessibility, fidelity [37], robustness [51], effectiveness, stability, and understandability [50]. However, there are no theoretical concerns that prompt favoring certain metrics over others [37]. Moreover, the literature lacks systematic approaches for assessing different explainability models in a comprehensive and balanced manner [51].

Consequently, the explainability of the RPE framework depends on evaluating the consistency between the visualizations produced by SHAP and LIME. Moreover, the SHAP and LIME values for the different words in the dataset were compared to prove the regularity of the impact of those words on the predicted rating class.

VII. RESULTS

The results of the transformer-based model experiments and the experiments conducted on the explainability models are presented in this section.

A. Rating Prediction Model Results

The results of applying the transformer-based models to both the English and Arabic datasets are presented in Table IX and Fig. 6. It shows that ELECTRA and AraBERTv2 achieved the best performance with respect to accuracy and F1 score.

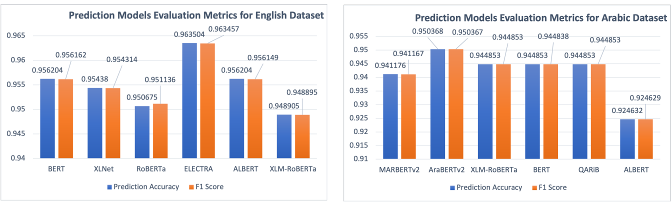


Fig. 6. Prediction performance.

In addition, the wilcoxon signed-rank test was applied to compare the performance of ELECTRA and AraBERTv2 against other baseline models in the same evaluation folds. The tests were conducted using paired accuracy and F_1 score obtained from identical evaluation settings. The results indicate that these two models achieved significant improvements over the other baseline models with $p < 0.05$ for accuracy and F_1 score. This confirms the effectiveness of the proposed transformer-based models configurations. Table X presents the statistical significance of test results.

B. Explainability Model Results

The transformer-based models that achieved the best performance, ELECTRA, and AraBERTv2, were used with the explainability models SHAP and LIME to justify the resulting predictions. Understanding the decisions made by a model is relevant for assessing the consistency of the resulting predictions and detecting possible biases in the model [38].

Two input sequences were used to visualize the explanations from each dataset. The selected reviews, along with their actual and predicted ratings, are listed in Table XI.

1) SHAP: A representation of the visualized explanations for reviews 1 and 2 from the English dataset is shown in Fig. 7. For review 1, all the words contributed to the prediction of rating class 1, with a total contribution of 0.46. For review 2, all the words contributed to the prediction of rating class 5, with a total contribution of 0.26, except for the words “to” and “and”, which contributed negatively, with a total contribution of 0.15.

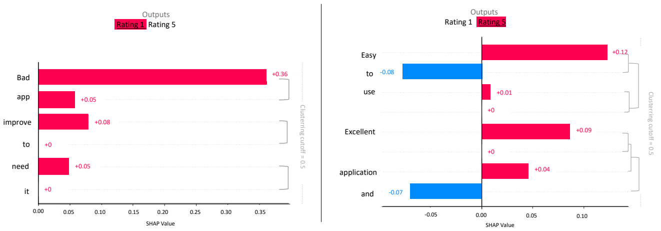


Fig. 7. SHAP visualization for English reviews where the Red bars mean positive contribution towards the predicted output and blue bars mean negative contribution towards the predicted output.

The visualized explanations for reviews 3 and 4 from the Arabic dataset are shown in Fig. 8. For review 3, the words “فاشل”, which means “failure”; “خطأ”, which means “error”; and “تعنيه”, which means “means”, contributed positively to the prediction of rating class 1, with a total contribution of 0.38, whereas all the other words contributed negatively to the predicted rating, with a total contribution of 0.20. For review 4, all the words contributed positively to the prediction of rating class 5, with a total contribution of 0.64, except for “وسيء”, which means “and bad”, which contributed negatively, with a total contribution of 0.14.

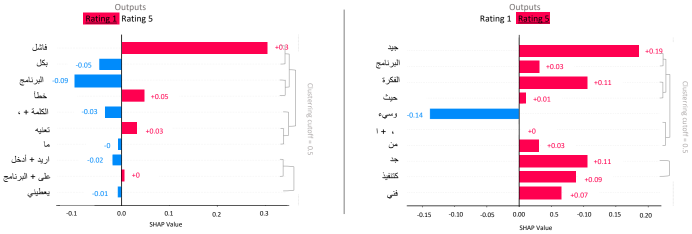


Fig. 8. SHAP visualization for Arabic Reviews where the Red bars mean positive contribution towards the predicted output and blue bars mean negative contribution towards the predicted output.

2) LIME: Each rating class in the LIME visualization is displayed with the probability of being the predicted class for a given input. For review 1 from the English dataset, the highest probability of 0.98 was for rating class 1. This probability occurred because of the contribution of all the words toward rating class 1, with a total of 0.75, except for the word “to”, which contributed positively to the prediction of rating class 5. On the other hand, review 2 from the same dataset had the highest probability of being of rating class 5, with a value of 0.99. All the words in the given review contributed to the prediction of rating class 5, with a total contribution of 0.33. Fig. 9 shows the LIME visualizations of these reviews.

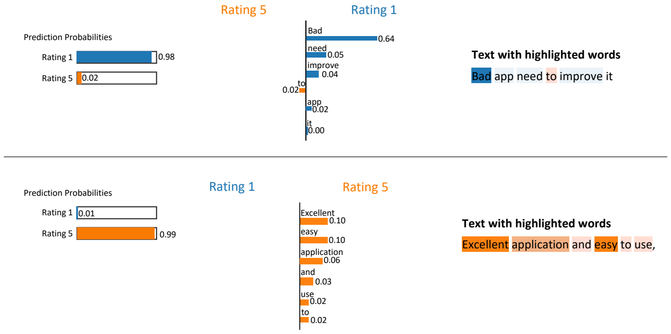


Fig. 9. LIME visualization for English Reviews where the orange bars represent the contribution towards rating 5 and the blue bars represent the contribution towards rating 1.

For the Arabic dataset, as shown in Fig. 10, review 3 had the highest probability of being of rating class 1, with a value of 0.91. The words that contributed positively to this rating class were “فاشل”, which means “failure”; “خطأ“, which means “error”; “تعنيه”, which means “mean”; and “ما“, which means “what,” with a total contribution of 0.40. All other words contributed negatively to the predicted class rating. Furthermore, review 4 achieved the highest probability, with

TABLE IX. Rating Prediction Results from Applying Different Transformer-Based Models

Dataset	Model	Accuracy	F_1	Precision	Recall
English	BERT	0.956204	0.956162	0.956530	0.956204
	XLNet	0.954380	0.954314	0.955024	0.954380
	RoBERTa	0.950675	0.951136	0.951136	0.950730
	ELECTRA	**0.963504**	**0.963457**	**0.964057**	**0.963504**
	ALBERT	0.956204	0.956149	0.956734	0.956204
	XLM-RoBERTa	0.948905	0.948895	0.948910	0.948905
Arabic	BERT	0.944838	0.944838	0.945462	0.944853
	MARBERTv2	0.941176	0.941167	0.941387	0.941176
	AraBERTv2	**0.950368**	**0.950367**	**0.950373**	**0.950368**
	XLM-RoBERTa	0.944853	0.944853	0.944853	0.944853
	ALBERT	0.924632	0.924629	0.924681	0.924632
	QARiB	0.944853	0.944853	0.944853	0.944853

TABLE X. Statistical Significance of Test Results

Model Comparison	Test	p-value
ELECTRA vs BERT	Wilcoxon	< 0.05
AraBERTv2 vs MARBERTv2	Wilcoxon	< 0.05

TABLE XI. Testing Reviews for Explainability Models

#	Review	Actual Rating	Predicted Rating
1	Bad application need to improve it	1	1
2	Excellent application and Easy to use	5	5
3	البرنامج فاشل بكل ما تعنيه الكلمة ، اريد أدخل على البرنامج يعطيني خطأ	1	1
4	البرنامج جيد من حيث الفكرة، وسيء جدا كتنفيذ فني	5	5

a value of 0.98 for rating class 5. The words “جيد”, which means “good”; “كتنفيذ”, which means “as execution”; “جدا”, which means “very”; “حيث”, which means “whereat”; and “من”, which means “from,” are the words that contributed positively toward the prediction of rating class 5, with a total contribution of 0.47. Conversely, the remaining words contributed negatively to the prediction of the rating class.

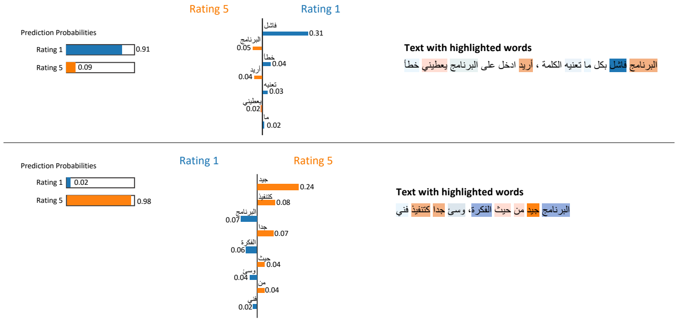


Fig. 10. LIME visualization for Arabic Reviews where the orange bars represent the contribution towards rating 5 and the blue bars represent the contribution towards rating 1.

In addition, as a proof of consistency between the explainability models used in the RPE framework, the average SHAP and LIME values were calculated for all the words that appeared in the reviews for apps related to social services for citizens/residents. The top ten words affecting each rating class are presented in Table XII and Table XIII.

TABLE XII. Top 10 Words Impacted Rating Class 1 from Reviews of Social Services for Citizens/Residents Apps

Dataset	Word	SHAP Value	Word	LIME Value
English	**crashing**	0.34	**slowly**	0.75
	useless	0.30	awful	0.59
	slowly	0.19	stop	0.57
	outdated	0.16	**outdated**	0.43
	charges	0.15	denied	0.42
	negative	0.15	cannot	0.35
	errors	0.11	asking	0.34
	frequently	0.11	pathetic	0.34
	feedback	0.09	hangs	0.33
	sorry	0.09	**crashing**	0.32
Arabic	**نصب (fraud)**	0.67	**نصب (fraud)**	0.56
	يفشل (fail)	0.66	**ضعيف (weak)**	0.34
	سيء (bad)	0.56	**سيء (bad)**	0.32
	بطيء (slow)	0.51	أزعجنا (annoyed us)	0.23
	أسوأ (the worst)	0.42	ينقص (lack)	0.23
	أحاول (trying)	0.42	اعتراض (objection)	0.21
	استخدام (usage)	0.41	يفقد (loose)	0.21
	أخطاء (errors)	0.39	الاستعمال (usage)	0.19
	ضعيف (weak)	0.38	شكاوي (complaints)	0.19
	فاشل (failure)	0.36	لا تعمل (do not work)	0.17

As seen in the tables, across both rating classes (1 and 5) and both languages (English and Arabic) SHAP and LIME show moderate-to-high consistency in identifying influential words, but with systematic differences in emphasis and ranking. SHAP tends to highlight words with strong global contribution to the model's prediction. Whereas LIME emphasizes locally influential features which may sometimes introduce context-specific words that do not appear among SHAP. This difference is expected given that SHAP is based on additive feature attribution, and LIME approximates the decision boundary locally around a single instance.

To measure how often both models identify the same words, Top-K Overlap Ratio is calculated as follows [see Eq. (5)]:

TABLE XIII. Top 10 Words Impacted Rating Class 5 from Reviews of Social Services for Citizens/Residents Apps

Dataset	Word	SHAP Value	Word	LIME Value
English	responsive	0.70	**worth**	0.65
	worth	0.63	enhancements	0.51
	great	0.55	solved	0.50
	nice	0.54	saves	0.43
	excellent	0.53	unbelievable	0.37
	reliable	0.53	accessible	0.34
	amazing	0.50	favourite	0.31
	well	0.49	finish	0.30
	valuable	0.48	**great**	0.30
	perfect	0.46	**perfect**	0.29
Arabic	كويس **(good)**	0.81	روعة (magnificent)	0.91
	إبداع **(creativity)**	0.40	كويس **(good)**	0.87
	ساعدني **(helped me)**	0.36	ساعدني **(helped me)**	0.67
	يعطيك (gives you)	0.35	رائعين (amazing)	0.59
	جميل (beautiful)	0.26	إبداع **(creativity)**	0.45
	سهل (easy)	0.26	حلو **(beautiful)**	0.41
	أفضل (best)	0.26	مميز (distinct)	0.41
	متطور (advanced)	0.23	تحسين (improving)	0.39
	دائمًا (always)	0.23	قوية (strong)	0.35
	عالمي (global)	0.12	ساعد (helped)	0.20

$$\text{Overlap@}K = \frac{\left|W^K_{\text{SHAP}} \cap W^K_{\text{LIME}}\right|}{K} \tag{5}$$

where, W^K_{SHAP} is the top-K words according to SHAP and W^K_{LIME} is the top-K words according to LIME.
Since each table reports top 10 words, K=10.

Rating Class 1: Common words in English are violation, abuse, and invading [see Eq. (6)]:

$$\text{Overlap@}10 = \frac{3}{10} = 0.30 \tag{6}$$

This means that there is a 30% overlap between both models [see Eq. (7)].

As for the Arabic, the common words are اتهاك, مزعج, بطيء, اختراق and, خطير

$$\text{Overlap@}10 = \frac{5}{10} = 0.50 \tag{7}$$

This means that there is a higher agreement in the Arabic reviews where there is a 50% overlap between both models.

Rating Class 5: Common words in English are perfect, useful, smooth, assist, served, and helpful [see Eq. (8)]:

$$\text{Overlap@}10 = \frac{6}{10} = 0.60 \tag{8}$$

This means that there is a 60% overlap between both models. This iddicates that both SHAP and LIME consistenly identify usability and performance as key drivers for high ratings.

As for the Arabic, the common words are فخر, سهل, and مميزة, غني

$$\text{Overlap@}10 = \frac{4}{10} = 0.40 \tag{9}$$

This means that there is a moderate agreement in the Arabic reviews where there is a 40% overlap between both models [see Eq. (9)].

Moreover, to help the app developers assess their apps and improve them based on the users' needs, they need to test the models on reasonable amounts of their own data (app reviews) to obtain a clear view of the features that affect users' opinions most strongly. In this study, the RPE framework was tested via reviews from the "Tawakalna" app, which belongs to the health sector and is an m-government app developed in Saudi Arabia. Table XIV and Table XV show the top ten words affecting each rating class.

TABLE XIV. Top 10 Words Impacted Rating Class 1 in "Tawakkalna" App

Dataset	Word	SHAP Value	Word	LIME Value
English	**violation**	0.92	**abuse**	0.64
	invading	0.90	hack	0.58
	complicated	0.85	restricted	0.53
	abuse	0.81	**invading**	0.51
	destroy	0.58	spying	0.38
	hanging	0.58	**violation**	0.30
	disgusting	0.50	steals	0.24
	lame	0.40	illegal	0.24
	crashes	0.40	invasion	0.23
	batteries	0.36	slow	0.21
Arabic	ثقيل (heavy)	0.94	يتأخر **(delayed)**	0.62
	اتهاك **(violation)**	0.60	اتهاك **(violation)**	0.46
	مزعج **(annoying)**	0.58	مزعج **(annoying)**	0.45
	تجسس **(spy)**	0.34	اختراق **(breakthrough)**	0.40
	خبيث **(malig- nant)**	0.32	خبيث **(malignant)**	0.36
	يتأخر **(delayed)**	0.28	خطير **(dangerous)**	0.35
	بطيء **(slow)**	0.25	بطيء **(slow)**	0.32
	اختراق **(breakthrough)**	0.25	يجبرك (forces)	0.31
	معلق (hanging)	0.23	صعب (hard)	0.28
	خطير **(dangerous)**	0.17	تجسس **(spy)**	0.27

VIII. Discussion

The goal of the RPE framework is to predict a rating for each review and justify the prediction using explainability models. However, the RPE framework focused on predicting the ratings for the reviews associated with rating classes 1 and 5 only. This came as a result of the non-benchmark data used in the study where there was a lack of reviews of rating classes 2, 3, and 4. The reasons behind that are reasons listed in Section V-D. Moreover, similar words were used in reviews with different rating classes; for example, the word "good" was used in reviews associated with the rating classes 3, 4, and 5; the words "problem," "fix," and "bad"

TABLE XV. Top 10 Words Impacted Rating Class 5 in "Tawakkalna" App

Dataset	Word	SHAP Value	Word	LIME Value
English	**perfect**	0.25	**useful**	0.42
	privacy	0.22	**perfect**	0.26
	useful	0.17	**smooth**	0.26
	served	0.17	**assist**	0.21
	smooth	0.16	easy	0.18
	rewarded	0.15	solution	0.18
	helpful	0.14	compatible	0.16
	assist	0.13	**served**	0.15
	intelligent	0.13	helps	0.15
	guide	0.08	effective	0.13
Arabic	مفخرة **(pride)**	0.28	متميز **(distinct)**	0.62
	متميز **(distinct)**	0.23	مفخرة **(pride)**	0.45
	جبار (huge)	0.18	غني **(rich)**	0.33
	غني **(rich)**	0.12	الخصوصيات (privacies)	0.21
	عالمي (global)	0.12	ساعد (helped)	0.20
	سهل **(easy)**	0.05	حلوا (solved)	0.19
	متقن (perfect)	0.04	سهلت (made easy)	0.17
	تطور (development)	0.04	اختصر (cut short)	0.16
	قوي (strong)	0.03	سهل **(easy)**	0.12
	احترافي (professional)	0.03	فريدة (unique)	0.12

were used in reviews associated with rating classes 1 and 2; "problem" was also used in reviews associated with rating class 3; and "fix" was used in reviews associated with rating class 4. Furthermore, the words "ممتاز" which means "excellent," and "رائع", which means "amazing," were used in reviews associated with rating classes 4 and 5, although "ممتاز" was used in reviews associated with rating class 3 as well. The words "للأسف", which means "sadly", and "المشكلة", which means "the problem", were used in in reviews associated with rating classes 1 and 2, whereas "المشكلة", which means "the problem", was also used in reviews associated with rating classes 3 and 4.

As a result, the RPE framework focused on predicting the ratings for reviews associated with rating classes 1 and 5 only. ELECTRA achieved the highest performance among other transformer-based models due to its efficient pre-training task known as Replaced Token Detection (RTD). This task works by masking a small percentage of tokens and trains on every token in the sentence. This allows the model to understand the context of the text more deeply with less computational power. On the other hand, AraBERTv2 achieved the best performance among the transformer-based models that works on Arabic language due to its specialized pre-training on a huge, diverse, and clean Arabic datasets. In addition to its ability to handle complex Arabic morphology. It excels in capturing semantic relationships between words, which is critical for rating predictions. As for the explainability, all the words that had either a positive or negative impact on a certain rating class based on SHAP value had the same impact on the same rating class based on LIME value as well with different impact. For example, the words "slowly", "outdated", and "crashing" from the English dataset and the words "نصب" which means "fraud", "ضعيف" which means "weak", and "سيء" which means "bad", from the Arabic dataset all had positive impacts on rating class 1, with different impact based on SHAP and LIME values. In addition, the words "worth", "great", and "perfect" from the English dataset and the words "كويس" which means "good", "ساعدني" which means "helped me", "إبداع" which means "creativity", and "حلو" which means "beautiful", all had positive impacts on rating class 5 with different SHAP and LIME values. This means that there is consistency between SHAP and LIME in explaining the resulting predictions.

However, to help app developers understand users' needs to improve their apps, the RPE framework should be applied to a reasonable number of reviews to identify the most important features affecting the apps. In this study, there was a lack of reviews for each app; hence, they could not be used alone to train the model. Accordingly, reviews for multiple apps were grouped to build the dataset.

To test the model, reviews for the "Tawakalna" app were used to apply the RPE framework. The results revealed the words that affected the resulting predictions the most, which can help the app developers improve the app on the basis of the users' perspectives. However, the computational cost for SHAP and LIME are very expensive. SHAP took around 16 hours to execute the code with an average of 1.92 minutes per record, whereas LIME took around 5 hours with an average of 0.6 minutes per record. This would significantly reduce the effectiveness of using explainability models for real-time explanations.

IX. Conclusion

The development of a mobile app review rating prediction model is fundamental for decision-making with respect to mobile app development. However, its predictions cannot be considered useful without a clear explanation. Explainability is effective for creating a foundation for trusting the results and helping decision-makers take action on the basis of these results. In this study, RPE framework is proposed in which it helps the decision-makers, app developers, and policy makers to evaluate their given apps and its usefulness based on the highly frequent words used by the users in their reviews in relation to the given ratings. In RPE, the rating prediction task is implemented using several transformer-based models. These models were applied to two separate datasets: English and Arabic reviews. The models help overcome the difficulties of not having an enormous amount of data to work with to obtain highly accurate results and of dealing with mixed-language datasets. The datasets were composed of reviews of Saudi m-government apps. In addition, explainability models were implemented to explain the resulting predictions using a visualization technique.

ELECTRA and AraBERTv2 achieved the best rating prediction performance among all other models with respect to accuracy and F1 score, while the SHAP and LIME explainability models justified the resulting predictions. Moreover, the RPE framework showed consistency between the explainability models used, where most of the time the same words contributed positively to the resulting predictions according to SHAP and LIME.

X. Study Limitations and Future Work

One of the main limitations of this study is the quality of the text used in the reviews. Data were collected specifically for this study and do not represent benchmark data; Much work needs to be done on the data to be able to collaborate with it. To address this limitation, additional data need to be used to obtain a wider variety of words to obtain a clear view of those that affect the resulting predictions. Moreover, the data sets were highly unbalanced with respect to the distribution among the rating classes. There was not enough reviews associated with rating classes 2,3, and 4. There are several reasons behind that as listed in Section V-D. This caused the study to focus only on the reviews associated with class 1 and class 5. Additionally, there is no general or bilingual transformer-based model to address different languages at the same time. As a result, the data collection, training, prediction, and explanation was done for each language separately. Further, to the best of our knowledge, there is a lack of studies in the field of m-government apps, especially for review rating prediction. Finally, the computational cost was remarkably high, especially with the consistency analysis that provided the SHAP and LIME average values for all the words in the reviews. This came as the result of the game theory that SHAP uses to provide explainability for a certain result. On the other hand, LIME's high cost comes from the many local model evaluations to fit surrogate models.

In future work, we plan to use prompt engineering as the initial context for the transformer-based models to guide the model's attention and influence its output. This will help the model understand the task and generate the desired output. Also, attention-based methods will be used to explain the resulting predictions. However, discriminating between the positive and negative contributions of the attention weights towards the predicted ratings must be considered during implementation. Furthermore, when it comes to m-government, other types of m-government apps should be explored to gain more knowledge about the perspectives of users about these apps in Saudi Arabia as a whole. In addition, applying this study to the m-government apps of other countries is necessary to improve the development of apps according to user needs. This could lead to the development of benchmark data for an international dataset of m-government apps that could serve all studies in this area. Further, user-centered evaluation for the proposed framework will be conducted by generating predictions and explainability for specific apps and check whether the apps' stakeholders and developers are benefiting from the provided results. Moreover, this work could be enhanced by the advancements in the multilingual transformer-based models. This could eliminate the duplication in doing the same work for both languages.

Data Availability Statement

The original data presented in this study are openly available in figShare at DOI: 10.6084/m9.figshare.25523716

Funding

No funding was received for this research.

Author Contribution

Conceptualization, Dhefaf Radain and Dimah Alahmadi; Data curation and Formal analysis, Dhefaf Radain; Methodology, Dhefaf Radain, Dimah Alahmadi and Arwa Wali; Project administration, Dimah Alahmadi and Arwa Wali; Writing - original draft, Dhefaf Radain; Writing – review and editing, Dhefaf Radain, Dimah Alahmadi and Arwa Wali.

References

[1] J. Chambua and Z. Niu, "Review text based rating prediction approaches: preference knowledge learning, representation and utilization," *Artificial Intelligence Review*, vol. 54, pp. 1171–1200, 2021.

[2] A. Rafay, M. Suleman, and A. Alim, "Robust review rating prediction model based on machine and deep learning: Yelp dataset," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. IEEE, 2020, pp. 8138–8143.

[3] S. Gheewala, S. Xu, S. Yeom, and S. Maqsood, "Exploiting deep transformer models in textual review based recommender systems," *Expert Systems with Applications*, vol. 235, p. 121120, 2024.

[4] Z. Y. Khan, Z. Niu, S. Sandiwarno, and R. Prince, "Deep learning techniques for rating prediction: a survey of the state-of-the-art," *Artificial Intelligence Review*, vol. 54, pp. 95–135, 2021.

[5] Z. Qiao, A. Wang, A. Abrahams, and W. Fan, "Deep learning-based user feedback classification in mobile app reviews," 2020.

[6] J. Dabrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: a systematic literature review," *Empirical Software Engineering*, vol. 27, no. 2, p. 43, 2022.

[7] H. O. Al-Sakran and M. A. Alsudairi, "Usability and accessibility assessment of saudi arabia mobile e-government websites," *IEEE Access*, vol. 9, pp. 48 254–48 275, 2021.

[8] W. Zhang, W. Gu, C. Gao, and M. R. Lyu, "A transformer-based approach for improving app review response generation," *Software: Practice and Experience*, vol. 53, no. 2, pp. 438–454, 2023.

[9] Z. Liu, "Yelp review rating prediction: Machine learning and deep learning models," *arXiv preprint arXiv:2012.06690*, 2020.

[10] A. A. Abdullah, S. H. Abdulla, D. M. Toufiq, H. S. Maghdid, T. A. Rashid, P. F. Farho, S. S. Sabr, A. H. Taher, D. S. Hamad, H. Veisi *et al.*, "Ner-roberta: Fine-tuning roberta for named entity recognition (ner) within low-resource languages," *arXiv preprint arXiv:2412.15252*, 2024.

[11] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based nlp," *AI*, vol. 4, no. 1, pp. 54–110, 2023.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] S. Das, N. Agarwal, D. Venugopal, F. T. Sheldon, and S. Shiva, "Taxonomy and survey of interpretable machine learning method," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 670–677.

[14] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[15] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42 200–42 216, 2020.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[17] Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost," *Computers, Environment and Urban Systems*, vol. 96, p. 101845, 2022.

[18] V. Korolev and P. Protsenko, "Accurate, interpretable predictions of materials properties within transformer language models," *Patterns*, vol. 4, no. 10, 2023.

[19] E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak, and M. Robnik-Šikonja, "Bert meets shapley: Extending shap explanations to transformer-based classifiers," in *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 2021, pp. 16–21.

[20] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153–4164, 2022.

[21] H. R. LekshmiAmmal, M. Ravikiran, and A. K. Madasamy, "Nitk-it_nlp@ tamilnlp-acl2022: Transformer based model for offensive span identification in tamil," *DravidianLangTech 2022*, p. 75, 2022.

[22] M. Alzahrani, A. Murshed, and M. Khayyat, "The role of m-government application in the saudi health sector in light of the covid-19 pandemic: A review," *TEM Journal*, vol. 11, no. 2, p. 731, 2022.

[23] Y. Al Moaiad, M. Alobed, M. Alsakhnini, and A. M. Momani, "Challenges in natural arabic language processing," *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 4700–4705, 2024.

[24] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[25] H. Yu, T. Qian, Y. Liang, and B. Liu, "Agtr: adversarial generation of target review for rating prediction," *Data Science and Engineering*, vol. 5, pp. 346–359, 2020.

[26] B. Wang, S. Xiong, Y. Huang, and X. Li, "Review rating prediction based on user context and product context," *Applied Sciences*, vol. 8, no. 10, p. 1849, 2018.

[27] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 364–373, 2018.

[28] B. H. Ahmed and A. S. Ghabayen, "Review rating prediction framework using deep learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 7, pp. 3423–3432, 2022.

[29] N. Aslam, W. Y. Ramay, K. Xia, and N. Sarwar, "Convolutional neural network based classification of app reviews," *IEEE Access*, vol. 8, pp. 185 619–185 628, 2020.

[30] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Discrepancy detection between actual user reviews and numeric ratings of google app store using deep learning," *Expert Systems with Applications*, vol. 181, p. 115111, 2021.

[31] I. Duru and A. S. Sunar, "Transformer and pre-transformer model-based sentiment prediction with various embeddings: A case study on amazon reviews," *Entropy*, vol. 27, no. 12, p. 1202, 2025.

[32] O. Orlovskiy, K. Sohrab, S. Ostapov, K. Hazdyuk, and L. Shumylyak, "Multilingual text classifier using pre-trained universal sentence encoder model," *Radio Electronics, Computer Science, Control*, no. 3, pp. 102–102, 2022.

[33] K. Kaur and P. Kaur, "Bert-rcnn: an automatic classification of app reviews using transfer learning based rcnn deep model," 2023.

[34] A. Shiju and Z. He, "Classifying drug ratings using user reviews with transformer-based language models," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2022, pp. 163–169.

[35] S. A. Chowdhury, A. Abdelali, K. Darwish, J. Soon-Gyo, J. Salminen, and B. J. Jansen, "Improving arabic text categorization using transformer training diversification," in *Proceedings of the fifth arabic natural language processing workshop*, 2020, pp. 226–236.

[36] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," *arXiv preprint arXiv:2004.06100*, 2020.

[37] T. Speith, "How to evaluate explainability?-a case for three criteria," in *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2022, pp. 92–97.

[38] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," *Journal of computer-aided molecular design*, vol. 34, pp. 1013–1026, 2020.

[39] M. Ahmed, M. S. Hossain, R. U. Islam, and K. Andersson, "Explainable text classification model for covid-19 fake news detection," *Journal of Internet Services and Information Security (JISIS)*, vol. 12, no. 2, pp. 51–69, 2022.

[40] S. Rao, Y. Li, R. Ramakrishnan, A. Hassaine, D. Canoy, J. Cleland, T. Lukasiewicz, G. Salimi-Khorshidi, and K. Rahimi, "An explainable transformer-based deep learning model for the prediction of incident heart failure," *ieee journal of biomedical and health informatics*, vol. 26, no. 7, pp. 3362–3372, 2022.

[41] K. Korangi, C. Mues, and C. Bravo, "A transformer-based model for default prediction in mid-cap corporate markets," *European Journal of Operational Research*, vol. 308, no. 1, pp. 306–320, 2023.

[42] A. Kurniasih and L. P. Manik, "On the role of text preprocessing in bert embedding-based dnns for classifying informal texts," *Neuron*, vol. 1024, no. 512, pp. 927–34, 2022.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[44] S. Wankmüller, "Neural transfer learning with transformers for social science text analysis," *arXiv preprint arXiv:2102.02111*, 2021.

[45] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *International conference on machine learning*. PMLR, 2020, pp. 9269–9278.

[46] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[47] A. Saranya and R. Subhashini, "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decision analytics journal*, p. 100230, 2023.

[48] K. M. Awlla, H. Veisi, and A. A. Abdullah, "Sentiment analysis in low-resource contexts: Bert's impact on central kurdish," *Language Resources and Evaluation*, pp. 1–31, 2025.

[49] R. M. Albalawi, A. T. Jamal, A. O. Khadidos, and A. M. Alhothali, "Multimodal arabic rumors detection," *IEEE Access*, vol. 11, pp. 9716–9730, 2023.

[50] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[51] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.