AUTOMATIC EVALUATION OF MOBILE HEALTH APPLICATIONS ACCORDING TO PERSUASIVE SYSTEM DESIGN PRINCIPLES AND MOBILE APPLICATION RATING SCALE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

YASİN AFŞİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

JUNE 2023

AUTOMATIC EVALUATION OF MOBILE HEALTH APPLICATIONS ACCORDING TO PERSUASIVE SYSTEM DESIGN PRINCIPLES AND MOBILE APPLICATION RATING SCALE

submitted by **YASİN AFŞİN** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Date: 20.06.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: YASİN AFŞİN

Signature :

ABSTRACT

AUTOMATIC EVALUATION OF MOBILE HEALTH APPLICATIONS ACCORDING TO PERSUASIVE SYSTEM DESIGN PRINCIPLES AND MOBILE APPLICATION RATING SCALE

AFŞİN, YASİN

M.S., Department of Information Systems

Supervisor: Prof. Dr. Tuğba Taşkaya Temizel

June 2023, 50 pages

Mobile applications have become an integral part of our daily lives and they have increasingly been used by the healthcare industry. However, many mobile health applications lack proper regulations or preliminary assessments, which could compromise users' health and safety. Existing literature has relied on manual assessment methods using Persuasive System Design (PSD) principles and Mobile Application Rating Scale (MARS) to evaluate user engagement and the quality of these applications. This thesis proposes a novel automatic evaluation technique to extend the assessment of applications in the market, with a focus on employing large language models to filter user reviews and generate sentence embeddings to classify the applications' use of PSD principles. Results show that it is possible to predict an application's employed PSD principles based on user reviews, while application descriptions do not provide enough information. Additionally, predicted classification probabilities of PSD principles are enriched with additional descriptive data, such as install counts and user ratings to predict MARS scores. Regression models trained using these methods outperform baseline models, with feature importance scores and SHAP values demonstrating the contribution of predicted classification probabilities of PSD principles to the models. Overall, this study suggests that automatic evaluation techniques can be effective in assessing the quality and user engagement of mobile health applications, providing an alternative to manual assessment.

Keywords: Data Mining, Mobile Health Applications, Persuasive System Design, Mobile Application Quality

ÖZ

MOBİL SAĞLIK UYGULAMALARININ İKNA EDİCİ SİSTEM TASARIM PRENSİPLERİNE VE MOBİL UYGULAMA DERECELENDİRME ÖLÇEĞİNE GÖRE OTOMATİK DEĞERLENDİRİLMESİ

AFŞİN, YASİN

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Prof. Dr. Tuğba Taşkaya Temizel

Haziran 2023, 50 sayfa

Mobil uygulamalar günlük hayatımızın ayrılmaz bir parçası haline geldi ve sağlık sektörü tarafından giderek daha fazla kullanılıyorlar. Ancak, birçok mobil sağlık uygulaması, kullanıcıların sağlığını ve güvenliğini tehlikeye atabilecek olmalarına rağmen uygun düzenlemelerden veya ön değerlendirmelerden yoksundur. Mevcut literatür, bu uygulamaların kullanıcı etkileşimlerini ve kalitelerini değerlendirmek için İkna Edici Sistem Tasarımı (PSD) ilkelerini ve Mobil Uygulama Derecelendirme Ölçeği'ni (MARS) kullanan manuel değerlendirme yöntemlerine dayanmaktadır. Bu tez, piyasadaki uygulamaların değerlendirilmesini genişletmek için özgün otomatik değerlendirme tekniği önerirken, kullanıcı yorumlarını filtrelemek ve cümle vektörlerini oluşturmak için büyük dil modellerini kullanarak uygulamalardaki PSD ilkelerini sınıflandırmaya odaklanmaktadır. Sonuçlar, uygulamalardaki PSD ilkelerini kullanıcı yorumlarına dayalı olarak tahmin etmenin mümkün olduğunu, uygulama açıklamalarının ise yeterli bilgi sağlamadığını göstermektedir. MARS puanlarını tahmin etmek için, PSD ilkelerinin tahmin edilen sınıflandırma olasılıkları, yükleme sayıları ve kullanıcı puanları gibi ek tanımlayıcı verilerle zenginleştirilmiştir. Eğitilen regresyon modelleri referans modellerden daha iyi performans sağlarken, öznitelik önem skorları ve SHAP değerleri, PSD ilkelerinin tahmin edilen sınıflandırma olasılıklarının modellere katkısını göstermektedir. Genel olarak, bu çalışma, otomatik değerlendirme tekniklerinin, mobil sağlık uygulamalarının kalitesini ve kullanıcı etkileşimini değerlendirmede etkili olabileceğini ve manuel değerlendirmeye bir alternatif sunabileceğini öne sürmektedir.

Anahtar Kelimeler: Veri Madenciliği, Mobil Sağlık Uygulamaları, İkna Edici Sistem Tasarımı, Mobil Uygulama Kalitesi

To Those Who Keep Going

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Tuğba Taşkaya Temizel, for her invaluable time, wisdom, guidance, support, patience, and generosity. Her contributions have been instrumental not only in the progress of my thesis research but also in shaping my overall academic trajectory. I consider myself fortunate to have received the support of her esteemed research team, who have provided valuable answers to my numerous inquiries.

Furthermore, I am deeply appreciative of the professors at the Informatics Institute of Middle East Technical University, whose teachings have profoundly influenced both my academic and professional endeavors. I would also like to extend my thanks to the members of the examining committee for their constructive feedback and insightful suggestions, which have greatly enriched my work.

I am deeply thankful to Dr. Tia Goss Sawhney for her unwavering support throughout this journey. Her guidance and encouragement have been invaluable to me.

I would like to thank all my family members, especially my mom and my sister, for their aid during the hard times that we all experienced.

I am grateful to our cat named 'Oğluş Anakin' for the commitment displayed during his transatlantic journey to continue being the backbone of our family and his contributions to this research.

Finally, I would like to express my heartfelt gratitude to my beloved wife, Seray Bircan, who has endured all the hardships with me. She is the source of all my energy during this period, and her presence is synonymous with joy and fulfillment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	Application Programming Interface
DD	Descriptions Dataset
LLM	Large Language Model
MARS	Mobile Application Rating Scale
MHA	Mobile Health Applications
MD	MARS Quality Scores Dataset
NLP	Natural Language Processing
PSD	Persuasive System Design
RD	Reviews Dataset
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

Mobile Health Applications (MHAs) are utilized by patients, healthcare providers, and caregivers for various purposes, such as keeping patient records, monitoring health conditions, educating about diseases, and providing motivation for specific diseases. Patients can easily access to these applications, which are a cost-effective way of providing healthcare services, as evidenced by various studies [1–3]. MHAs have even been investigated as prescribable non-pharmaceutical solutions to be offered by general practitioners to patients for specific diseases that require constant monitoring and management of health conditions [4].

The popularity of MHAs has resulted in the publication of a significant number of mobile applications, 41.5 thousand in Apple's App Store and 54.5 thousand in Google Play Store for the third quarter of 2022 [5, 6]. However, studies show that low-quality MHAs can cause harm to patients by providing incorrect information or inappropriate responses to their needs [7]. Therefore, mobile application evaluations for the effectiveness and quality of MHAs are necessary, as the market continues to expand with new applications.

In the literature, Persuasive System Design (PSD) [8] is a framework that describes principles and techniques for developing or assessing the abilities of mobile applications to influence user behaviors and attitudes. Additionally, the Mobile Application Rating Scale (MARS) [9] has been developed for the quality assessment of MHAs covering different aspects of the application, such as functionality, engagement, and aesthetics. However, both these frameworks rely on manual expert evaluations. These experts are selected from a range of healthcare related professionals, including medical doctors. Additionally, they should be trained about the specifications of each framework to achieve consistent evaluations, such as the meaning of each item in the questionnaire or examples of PSD principles for the specific MHA category. As a result, manual assessments are inefficient in terms of time and expert allocation, resulting in limited market prevalence compared to a large number of published mobile health applications.

In this thesis, automatic evaluation techniques are suggested as an alternative to manual assessments. Three datasets are created that contain application descriptions, user reviews, and application qualities. These datasets are filtered using textual entailment-based classification using pre-trained large language models. The filtered textual data is used for sentence embeddings, and they are enriched with descriptive mobile appli-

cations data. Using the feature sets, supervised machine learning models are created and used for testing the hypothesized ideas of automatizing PSD principle predictions and MARS quality scores of MHAs.

1.1 Research Questions

This thesis aims to answer the following research questions.

RQ1 How can we predict PSD principles of an expert-evaluated MHA based on user reviews and application descriptions?

It is hypothesized that collected user reviews and application descriptions contain the MHA's main features, functionalities, and purposes from the perspectives of users and developers. Hence, user reviews and application descriptions can be relevant sources for conducting content analysis regarding employed PSD principles of MHAs. To investigate this, large language models and binary classification models are utilized for PSD principles predictions.

RQ2 How can we predict MHA quality on the scale of MARS by using the collected descriptive MHA information? Can PSD principles that are predicted by the models contribute to this prediction?

It is hypothesized that collected descriptive data (e.g., overall customer rating, install counts, advertisement support) and predicted classification probabilities of PSD principles could be used as a significant feature vector to predict MARS overall scores. To explore this, regression models are developed, and features importance is calculated.

1.2 Contributions of the Study

The contributions of this study are as follows:

- Presenting MHA datasets and example approaches to collect and enrich them from multiple sources,
- Presenting example approaches to detect related textual data with the given PSD definitions,
- Proposing models to predict employed PSD principles of MHAs using user reviews and application descriptions,
- Proposing models to predict MARS quality scores of MHAs using descriptive MHA data and PSD principle predicted probabilities,
- Analyzing the contribution of PSD principles predicted probabilities on the regression of MARS quality scores,

1.3 Organization of the Thesis

In the thesis, Chapter 2 is about previous studies and the research gap that the thesis fit. Chapter 3 explains the implemented architecture and methodologies used in it. Chapter 4 explains dataset collection and their analysis, experiment results, and ablation studies. Chapter 5 summarizes and further discusses the results. Chapter 6 concludes the thesis and explains limitations and possible future works.

CHAPTER 2

RELATED WORK

In this chapter, a review of studies related to the thesis is conducted. These studies can be categorized into three subsections: persuasive system design, mobile application quality analysis, and language models. The previous studies are also examined to identify the research gap in this section.

2.1 Mobile Application Analysis

Mobile application reviews posted on mobile application platforms have been the focus of many studies in the literature. In these studies, user reviews are categorized into meaningful groups, such as feature requests, bug reporting, and information giving, as seen in [10], or using lower-level taxonomies, such as battery usage, security, application price, and security, as observed in [11]. Novel frameworks for user review analysis, such as AR-miner [12] and CLAP [13, 14], have been proposed to filter and prioritize informative reviews for application developers. Other studies, such as [15–18], have aimed to improve application release planning and feature elicitation by utilizing user reviews as a rich source of information.

Some studies have also analyzed the application descriptions that developers publish on mobile application pages. For example, in [19], application descriptions and user reviews are combined to identify key features that lead to higher user ratings. Similarly, [20] notes that descriptions provide a summary of the features implemented by developers, while reviews provide test comments on those features. Their approach has led to the development of frameworks for application feature health checks, feature recommendations, and feature optimizations. SimApp [21] is another study that shows that application descriptions and user reviews are effective in finding similar applications to generate user suggestions. Application descriptions can also serve as content summarizers for evaluating application content, as seen in [22, 23], where application maturity levels are predicted using extracted features from application descriptions. Maturity-level decisions require evaluation of mobile applications in various aspects, and the success of these studies shows that application descriptions can summarize the general content of mobile applications.

These studies show that user reviews and application descriptions can be the source for the various purposes of mobile application analysis. Further, studies of mobile applications that are related to mobile application qualities and persuasive system designs are explained in the following subsections.

2.1.1 Mobile Application Quality

Although mobile application regulations and rules are mandated by the platforms [24, 25], mobile application qualities can vary. User ratings, reviews, and the number of downloads are not reliable indicators of quality, especially for Mobile Health Applications (MHAs) [26]. Therefore, systematic evaluations of MHAs are important to eliminate potential adverse events [27]. In the literature, there are three tools available to assess the quality of mobile applications.

The first tool is Mobile Application Development and Assessment Guide (MAG), which uses eight criteria to assess the development and design processes, such as usability, privacy, security, suitability, transparency, safety, technical support, and technology [28]. The second tool is the IMS Institute for Healthcare Informatics mobile application functionality score, which evaluates the functionalities of applications using seven criteria [29]. The third and most dominant tool is the Mobile Application Rating Scale (MARS), which assesses mobile health application qualities using a five-point Likert scale in four main groups: engagement, functionality, esthetics, and information quality [9]. The MARS has been validated and translated into German and Italian [30, 31] for researchers to use in different languages. MARS's additional ability, compared to other scales, to assess the appropriateness and accuracy of the information in the application is also important for patient safety and health [32].

MARS has four groups of questions based on its questionnaire. These questions cover the dimensions that are shared below:

- *Engagement*: This dimension focuses on the user experience during the engagement of the mobile application. It covers entertainment, customization, interactivity and target group aspects of the mobile application.
- *Functionality*: This dimension focuses on the capabilities of the mobile application. It covers performance, ease of use, navigation, and gestural design aspects of the mobile application.
- *Aesthetics*: This dimension focuses on the visual design of the mobile application. It covers the layout, graphics, and visual appeal aspects of the mobile application.
- *Information quality*: This dimension focuses on the information content of the mobile application. It covers the application description accuracy, goals, information quality and quantity, visual information, credibility and evidence aspects of the mobile application.

Studies have investigated the relationship between MHA qualities in the scale of MARS and application specifications using correlation analysis. While some studies have checked the correlation between the overall MARS scores and the application's last update date, install counts, developer affiliation, and application categories [33–35], most studies focused on user ratings. Some studies could not find any significant correlation between MARS scores and user ratings [33–36], while others have found a significant correlation [37, 38].

Studies mentioned in this section either focus on quality evaluation tools or specifications that have a positive effect on the qualities. All these studies conduct their research with a focus on health applications, and this also indicates the additional need for quality analysis for the health applications compared to the other mobile application categories.

2.1.2 Persuasive System Design

Information technologies as persuasive systems affect user behaviors and actions [39]. In the Fogg Behavior Model, persuasion is explained with motivation, ability, and triggers drivers [40]. There are also other frameworks that study the influence strategies. Cialdini describes reciprocity, commitment and consistency, social proof, authority, liking, and scarcity principles of persuasion [41]. Behavioral Change Technique (BCT) taxonomy shares 26 principles for persuasion [42]. In the persuasive system design (PSD) framework, which is popular for mobile applications and suggested by Oinas-Kukkonen and Harjumaa, persuasion postulates, contexts, and, most importantly, 28 persuasive design principles are explained [8]. On contrary to BCT principles, PSD principles are grouped into four categories: primary task support, dialogue support, social support, and system credibility by giving theoretical examples [43]. While BCTs can be implemented in daily life routines without technologies, PSD principles are more custom and need technologies to persuade users [44, 45].

According to study of Oinas-Kukkonen and Harjumaa [8], 28 PSD principles are defined as follows:

- *Reduction*: Reducing the task complexities
- *Tunneling*: Guiding users with a series of experience
- *Tailoring*: Customising the content according to user groups
- *Personalization*: Changing the system for the liking of the user
- *Self-monitoring*: Enabling users to track their records
- *Simulation*: Enabling users to see the cause and effect of actions
- *Rehearsal*: Enabling users to practice a behavior
- *Praise*: Motivating user with praises

- *Rewards*: Providing incentives for the target behavior
- *Reminders*: Prompting notifications for the target behavior
- *Suggestion*: Providing fitting recommendations to users
- *Similarity*: Establishing similarity between the system and the user
- *Liking*: A system that visually appealing to users
- *Social role*: Offering ways that adapt a social role
- *Trustworthiness*: Building a user confidence towards the system
- *Expertise*: Offering information that is known as expert knowledge
- *Surface credibility*: A system that has a competent first look
- *Real-word feel*: Providing information about the backstage of the system
- *Authority*: Referencing to roles that have authority
- *Third-party endorsements*: Having endorsements from respected organizations
- *Verifiability*: Providing means to be used for the information verification
- *Social learning*: Providing means for observing other users
- *Social comparison*: Providing means for comparing other user performances
- *Normative influence*: Providing means for emphasizing the norms
- *Social facilitation*: Providing means for recognizing other users with the same behavior
- *Cooperation*: Providing means to users that they can work together
- *Competitions*: Providing means to users that they can compete with each other
- *Recognition*: Providing public recognition for performing the target behavior

Although the PSD model offers strategies for developing new applications, it can also be used to identify the principles used in already developed and published applications. PSD principles have been investigated for mental health applications [46–48] and personal well-being applications [49]. In [50], diabetes-related MHAs were analyzed, and employed PSD principles were detected through user reviews. Similarly, in [51], primary task support principles were investigated for mental health mobile applications using natural language processing (NLP) techniques and topic modeling. The study found that the results are similar to manual expert assessments of employed PSD principles [48]. Moreover, in [52], diabetes self-management MHAs were analyzed in terms of the relationship between application qualities, user ratings, and employed PSD principles. Application qualities were assessed using MARS for 120 MHAs by trained medical experts. These experts also evaluated MHAs' PSD

principles for each application by downloading and testing each one. All MHA assessments and examples for each PSD principle were shared as a data repository for future research. The study found that PSD principles correlate with application quality scores and improve user engagement with MHAs.

2.2 Natural Language Processing

In the field of artificial intelligence, computational language processing is studied under the subfield called Natural Language Processing (NLP) [53]. NLP is a popular research area that encompasses Natural Language Understanding (NLU) and Natural Language Generation (NLG) [54]. Initially, NLP relied on rule-based approaches, but with the advent of statistical techniques, large datasets could be utilized to identify patterns [55]. Recent advancements in deep learning models, such as LSTM [56], and CNN [57], have resulted in NLP achieving its state-of-art performances [58]. Large language models such as BERT [59], RoBERTa [60], T5 [61], XLNet [62], and GPT models [63] have been developed with the breakthrough of transformers [63]. These models are pre-trained on vast data and fine-tuned on specific tasks, leading to significant performance improvements. Various tasks, including text classification, machine translation, text generation, and question answering, have been shown to be effectively performed using NLP techniques [64]. Furthermore, NLP has been applied to health-related data, including extracting information about symptoms, medications [65–67], and potential medical problems [68]. Conversational agents like chatbots, powered by large language models, can aid patients by allowing them to ask questions for appropriate advice on their health conditions [69, 70].

In addition, NLP techniques have been used to automate expert evaluations in various fields. For instance, in [71], NLP features are generated to automatically classify stroke patients. In [72], an information extraction tool is utilized to automatically detect patients with congestive heart failure risk from medical patient records. Moreover, in [73], the NLP pipeline is reviewed to eliminate manual reviewing of expert explanations in radiology reports, while [74] demonstrates the importance of automated NLP-based error prevention tools to prevent human errors in patient screenings. Additionally, NLP assistance is even used in new tools to ease researcher annotation tasks for biomedical data [75]. NLP techniques enable conducting comprehensive analysis replacing manual expert evaluations while also eliminating human errors and time waste.

2.3 The Research Gap

In the literature review, it is observed that previous studies have explored the aspects of PSD principles and mobile application qualities, but they are either limited in scope or conducted with manual evaluations. The studies that have investigated PSD principles have relied on downloading and experiencing the applications, and no research has been conducted to detect PSD principles from textual data using automatic predic-

tion methodologies. Similarly, application qualities have been determined based on simple indicators such as user ratings, and studies that predict MARS quality scores have used linear regression-based methods, which do not account for nonlinear interactions. Therefore, there is a gap in the literature for the use of more advanced machine learning methods to predict the qualities of mobile health applications.

The study by Geirhos et al. [52] is a standardized evaluation study that considers both PSD principles and MARS quality scores of mobile health applications. This study, which serves as motivation for this thesis, reveals that there are correlations between four categories of PSD principles and MARS quality scores. Moreover, as stated in the study, a significant ratio of the quality score's variance is explained with dialogue support and social support categories. The study also shares MHA lists and expert annotations, making it possible to conduct automatic assessments of MHAs in terms of both PSD principles and MARS quality. Additionally, the predictability of MARS quality scores using PSD principles can contribute to the persuasiveness literature.

CHAPTER 3

METHOD

The thesis conducts experiments in four steps: dataset collection and preprocessing (Step 1), filtering relevant text (Step 2), PSD principles prediction (Step 3), and MARS quality score prediction (Step 4), as shown in Figure 1. In the following subsections, each step is detailly explained in the same order shown in the architecture diagram.

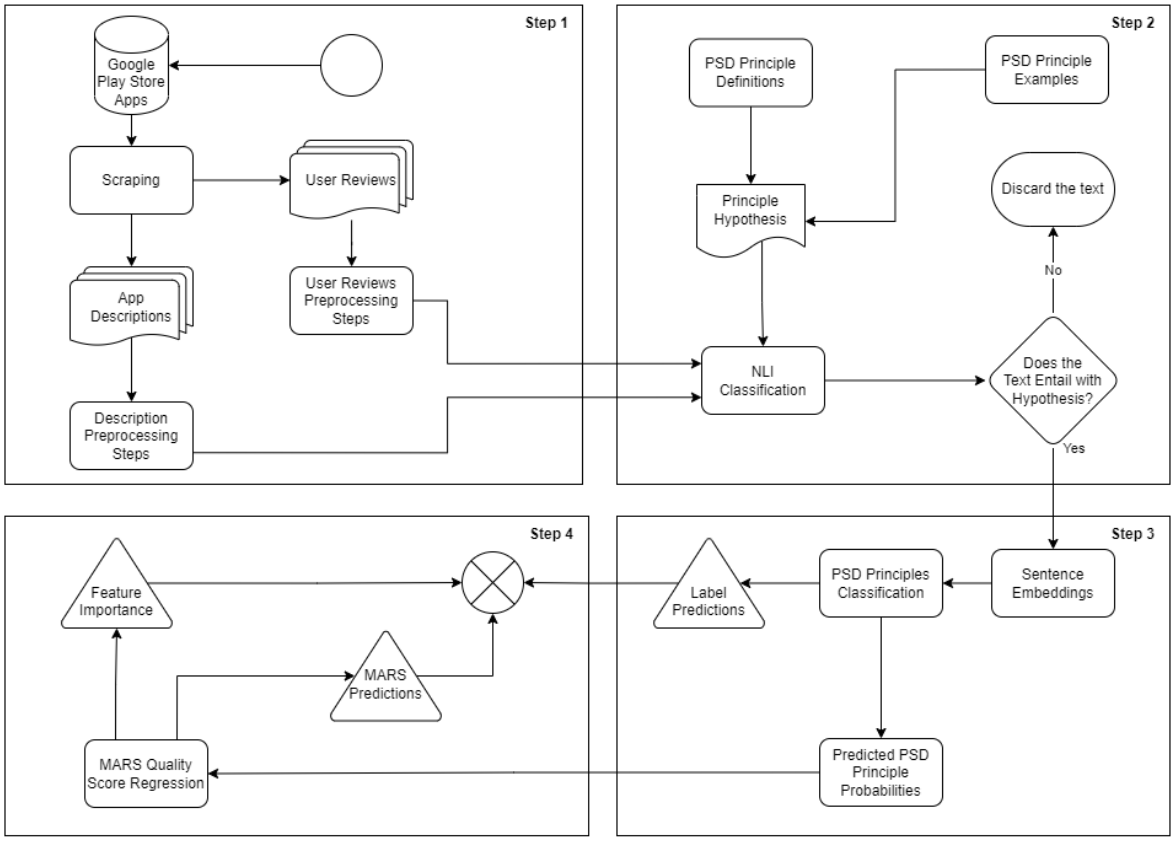

Figure 1: Architecture diagram

3.1 Dataset Collection and Preprocessing

The datasets are collected using the Google Play Store Scraper API [76] for both user reviews and application descriptions in this master thesis. The collection process includes all languages to ensure the acquisition of all available user reviews and non-English reviews are translated using the Google Translate API [77]. To prepare the data for analysis, preprocessing is applied to both user reviews and application descriptions to remove emojis, hashtags, and special characters. Application descriptions consist of itemized sentences without any punctuation, as seen in Figure 2, which requires the addition of dots at the end of each itemized sentence to differentiate them in raw text format. Additionally, reviews containing less than two words are eliminated, as longer sentences are required to effectively reference PSD principles.

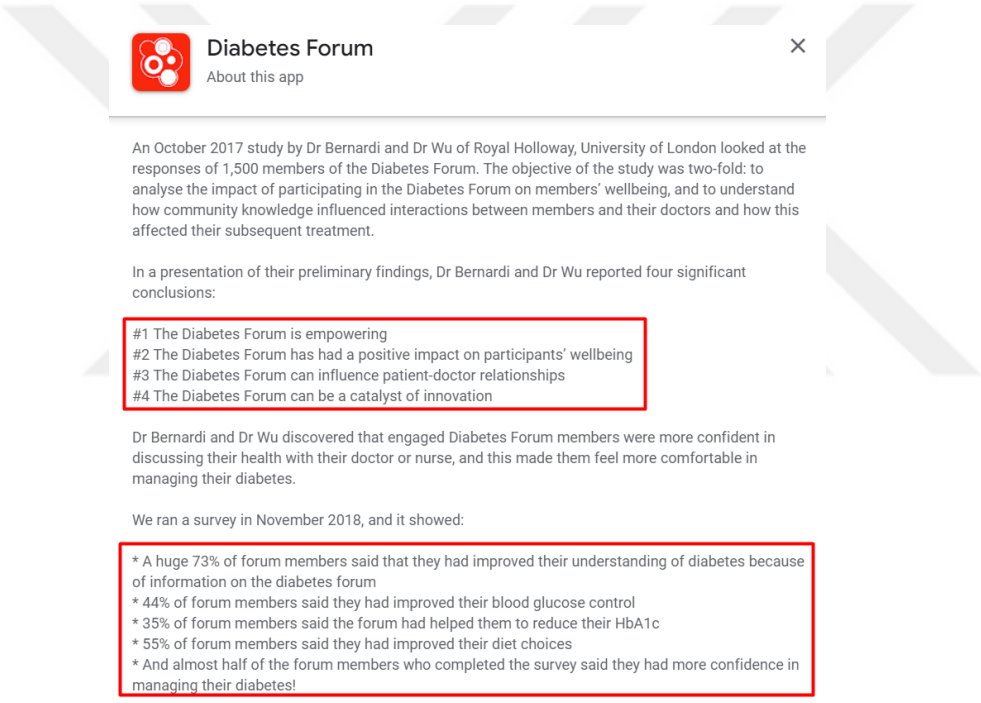

Figure 2: An application description example for the app 'Diabetes Forum'. The description contains itemized sentences, which require preprocessing to be analyzed in raw text format.

3.2 Textual Entailment

Each MHA in this study contains multiple user reviews, but not all of them are informative in terms of PSD principles. Similarly, while application descriptions serve to market the application, they may not necessarily be related to PSD principles. Therefore, in this step, Natural Language Inference (NLI) classification is implemented to

determine their relevance to PSD principles. In the NLI classifiers, two sentences, which are called premise and hypothesis, are used to evaluate their logical relations. While premises are assumed truths or conditions, hypotheses are tested with these assumptions in terms of their logical relations. NLI classifiers evaluate whether a given hypothesis and premise are entailed (support), contradicted (oppose), or neutral (not enough info). For a general context premise, hypothesis examples for each of the possible classes are shared in Table 1. As can be seen from the given examples, entailed hypothesis is a supporting sentence that is focused on different aspects when describing the cat in the room. While the neutral hypothesis is unrelated to the given premise, the contradicted hypothesis mentions the opposite behavior of the cats to the mentioned one in the premise.

Table 1: NLI classification examples for *Entailed*, *Contradicted* and *Neutral* hypotheses with a given premise.

Premise	A cute orange cat is sleeping on the sunny side of the couch.
Entails	A small cat is sleeping in the room, with a song in the background.
Neutral	A dog can be a loyal companion if you need emotional support.
Contradicts	Cats never sleep under the sun, and they mostly like cold places.

The quality of textual data is critical in this study, as supervised learning methods are being used in the subsequent PSD principle classification step. Thus, NLI classifications are implemented for both user reviews and application descriptions to identify their relativeness with PSD principles. To achieve NLI classifications, separate premises are generated from user reviews and descriptions. Moreover, hypotheses are generated for each PSD principle based on the descriptions provided by Oinas-Kukkonen and Harjumaa [8], and examples are provided to experts during PSD principle labeling [52]. Since the aim of this step is to filter the on-hand textual data, hypotheses are generated accordingly, as shared in Table 2. Each hypotheses sentence is created starting with *"The text is about..."*, so the direct semantic relation between the hypothesis and the input text is classified. While neutral classifications do not show enough information to be related, contradicted classifications are accepted against the given hypothesis, which means they are not about the PSD principles as well. Hence, only the inputs that are classified as entailed are selected for the following steps, and other texts are discarded. Additionally, an ablation study is conducted in Section 4.3.2 to test the effectiveness of numerous hypotheses for PSD principles with different sentence structures, as the determination of the hypothesis significantly affects the performance of the overall system.

For the textual entailment task, the XLNET [62] NLI classifier, which has been fine-tuned on SNLI [78], MultiNLI [79], and FEVER [80] datasets, is selected based on its performance in the Adversarial NLI (ANLI) benchmark [81]. An ablation study is also conducted and explained in Section 4.3.1 to examine the effects of selected textual entailment models on PSD principles' classification performance.

Table 2: Examples and generated NLI hypotheses for *Tailoring*, *Self-monitoring*, *Praise*, and *Reminders* principles.

	PSD Principle Examples	**Hypothesis**
Tailoring	• Adapting information according to type of diabetes. [52] • Personal trainer Web site provides different information content for different user groups, e.g. beginners and professionals. [8]	The text is about providing adaptive information for users.
Self-monitoring	• Feature for glucose tracking. [52] • Heart rate monitor presents a user's heart rate and the duration of the exercise. [8]	The text is about tracking user performances.
Praise	• Feedback/compliments on tracked data. [52] • Mobile application that aims at motivating teenagers to exercise uses text messages with praises. [8]	The text is about motivating users with praises.
Reminders	• System based daily pop-up messages. [52] • Caloric balance monitoring application sends text messages to its users as daily reminders. [8]	The text is about sending notifications to users.

3.3 PSD Principles Prediction

Hypothesis-related filtered user reviews for each MHA are concatenated and formed different paragraphs. Using the ST5-large sentence transformer model [82], sentence embeddings are generated from these paragraphs. The model is considered the highest-scoring model for paragraph-based text classification tasks by the Massive Text Embedding Benchmark (MTEB) [83]. ST5 family outperforms the second-best model by an average of 3% for eight text embedding tasks in this benchmark.

Generated text embeddings are used in machine learning classification models: LightGBM [84], Support Vector Machine (SVM) [85], and Elasticnet [86] classifiers to predict the employed PSD principles of MHAs. These models are selected for their ability to handle high data dimensionalities, which is a common challenge in text classification tasks. On the other hand, SVM and Elasticnet models handle overfit-

ting and high dimensionality with their regularization parameters. Although LightGBM models are prone to overfitting in small datasets, ensemble methods are used in the literature to overcome the problem [87, 88]. Overall, the combination of these three models provides a diverse set of approaches for predicting the PSD principles of MHAs using text embeddings as input.

The hyperparameters for the LightGBM model are optimized using the Optuna framework [89], while the hyperparameters for the SVM and Elasticnet models are selected using grid search. Given the small size of the dataset, the fold numbers for each implementation are kept high, equal to the count of positive labels for PSD principles. High fold numbers enable larger training splits, with each positive-labeled data also placed in a separate test fold. The trained models for each fold are saved with their data splits to be used for predicting unseen datasets. Predicted PSD principle classification probabilities for each MHA are calculated as many times as the number of fold numbers in this ensemble method.

The models are trained, validated, and tested using nested stratified k-fold cross-validation method. The Matthews Correlation Coefficient (MCC) is the primary performance metric used in binary classification, which considers all four confusion matrix categories and is particularly effective for imbalanced datasets [90]. Additionally, accuracy, precision, recall, and F1 scores are also reported.

3.4 MARS Quality Scores Prediction

In this step, to predict MARS quality scores of MHAs, categorical and continuous features, in addition to predicted classification probabilities of PSD principles, are used. These descriptive features are collected from the application store. Install counts are ordinal categorical features, while predicted probabilities and average user ratings are continuous. Additionally, the presence of advertisements, published privacy policies, shared developer addresses, advertisement support, and shared developer website are used as binary categorical features. Moreover, by definition, MARS quality scores are bound by maximum and minimum values and acquired quality scores in this study exhibit nonparametric behavior that does not fit known distributions.

Considering the mentioned mixed feature types and nonparametric behavior of the target variable, decision tree-based LightGBM [84] and CatBoost [91] regressor models are selected and trained using the 5-fold nested cross-validation method. CatBoost is preferred due to its encoding of categorical features, which is the predominant feature type of this step, and its use of ordered target statistics differently than the LightGBM model. Moreover, hyperparameter tuning of these models is achieved with the Optuna framework [89]. Model performances are calculated and compared with the baseline models, using the mean absolute error (MAE), root mean squared error (RMSE) and relative root mean squared error metrics (Relative RMSE). Relative RMSE is selected as the primary performance metric since it reports the error rate as in the percentage ratio of the mean observed values.

Furthermore, feature contributions are calculated with both finetuned and trained LightGBM and CatBoost regression models. Built-in feature importance functions of decision-tree-based algorithms are used to interpret feature importance in terms of information gains in the splits of decision trees. Additionally, SHAP values [92] are calculated to determine direct feature impacts on model outputs. Considering the results of these two methods, the least contributing features are detected, and regression models are again experimented excluding these features to evaluate if predictions improve with fewer features.

CHAPTER 4

EXPERIMENTS

4.1 Datasets

In this study, three datasets are created using two different data sources for each. The creation steps and data analysis for these datasets are as follows:

4.1.1 PSD Principles Datasets

In the study of Geirhos et al. [52], trained experts assessed 120 MHAs regarding 24 of the 28 PSD principles recommended by Oinas-Kukkonen and Harjumaa. The dialogue support category's liking principles and the system credibility support category's trustworthiness, expertise, and surface credibility principles were excluded as they are non-technical. The study's MHA lists consisted of application and developer names and binary PSD principle labels (true if they exist). The MHA list from the study contains only 31 applications from Apple's App Store. However, any of these could not be found via the web, mobile, or native API of the application store because Apple's App Store has a policy of removing applications if they are not recently updated or downloaded by users [93]. On the other hand, 75 of 89 listed Google Play Store applications were found by matching application and developer names. The unique application ids were extracted from the web URLs of these applications and used as search keys for Google Play Store Scrapper API [76].

Application descriptions, description summaries, and available user reviews were scraped on 02-May-2022. The Descriptions Dataset (DD) was created containing scraped application descriptions User reviews were only found for 45 of 75 MHAs, and the user reviews were discarded if they were posted after the data collection date of expert evaluations (December 2020) since these reviews were assumed to be posted for the updated versions of MHAs. On the other hand, description summaries were not included in dataset creations since they were found to be not informative compared to application descriptions and user reviews. For instance, description summary of *"Do you want to know about type 2 diabetes diet plan? Hit installs for secrets."* is a marketing phrase to promote the application in general instead of pointing out specific features that the study aims for.

In supervised learning, balanced datasets prevent models from being biased towards a single class. In the PSD datasets, principles with ratios closer to 50% are considered more balanced. The employed PSD principle (true label) ratios for all MHAs and user-reviewed ones are shared in Table 3. Considering these ratios, *Tailoring*, *Self-monitoring*, *Praise*, and *Reminders* principles are selected to conduct experiments from both the Reviews Dataset (RD) and Descriptions Dataset (DD) based on the employed principle ratios and the higher correlations between MARS Quality Scores and PSD categories of Primary Task Support and Dialogue Support found in Geirhos et al.'s study [52].

Table 3: Employed PSD principles' ratios of DD and RD

PSD Categories	**Principle Names**	**Implemented Principles (%)**	
		DD	**RD**
Primary Task Support	Reduction	13.2	17.8
	Tunneling	7.9	8.9
	Tailoring	**28.9**	**33.2**
	Personalization	10.5	13.3
	Self-monitoring	**30.2**	**37.7**
	Simulation	1.3	2.2
	Rehearsal	2.6	2.2
Dialogue Support	**Praise**	**28.9**	**35.6**
	Rewards	2.6	0.0
	Reminders	**21.1**	**33.3**
	Suggestions	86.8	80.0
	Similarity	3.9	4.4
	Social Role	7.9	8.9
Social Support	Social learning	3.9	2.2
	Cooperation	2.6	0.0
	Social facilitation	10.5	11.1
	Recognition	7.9	11.1
	Normative Influence	11.8	15.6
System Credibility	Real world feel	98.7	91.1
	Authority	21.1	17.8
	Third party endorsement	32.9	28.9
	Verifiability	26.3	24.4

4.1.2 MARS Quality Score Dataset

Mobile Health App Database (MHAD) is an open science repository for MARS quality ratings of MHAs [94]. Since MHAD is only published over a user-interactive web interface, a custom web scrapper is developed using Python's Selenium library to create an MHAs list. The list comprises 489 MHAs' names, webpage links, and MARS quality scores. On the data collection date (04-May-2022), 194 MHAs are found to have user reviews. In addition to user reviews, additional descriptive features are collected with Google Play Store API [76] for these MHAs. Collected variables are merged with the scraped MARS scores, and MARS Quality Scores Dataset (MD) is created (primary features). The details of MD fields are shown in Table 4.

4.2 Results

In the following subsections, PSD principles' predictions are separately experimented with regarding user reviews and descriptions to answer RQ1. Further, MARS quality scores are predicted, and the features importance of the predictions is calculated to answer RQ2.

4.2.1 PSD Principles Prediction From User Reviews

To the best of my knowledge, no study has predicted employed PSD principles in an automated way. Therefore, simple baseline models are used in this experiment to compare trained model performances. The first baseline model (B1) predicts labels randomly using the true label weights. The second baseline model (B2) predicts labels as the most frequently observed label in the training dataset to be used in accuracy comparisons. User reviews undergo textual entailment filtering with each target PSD principle's hypothesis, using the pre-trained XLNET NLI classifier. Table 5 presents user review examples and their textual entailment classifications. This step is implemented as zero-shot classification since there is no ground truth for the labels. As can be seen from the given examples, hypothesis-related reviews are correctly classified as *Entails*. However, there are interchangeably misclassified user reviews for the labels of *Neutral* and *Contradicts*. For instance, the user review of *"Easy to keep up with has reminders easier for older people to use"* is classified as *Contradicts*, although there is no evidence that contradicts with the hypothesis of *Praise* principle. Since both of the *Neutral* and *Contradicts* labels are assumed as "not related" with the used hypotheses and discarded from datasets, the misclassification between these labels does not affect the implementation in this step.

Table 4: Primary and secondary feature details of MD

Feature Source	Feature Name	Data Type	Remarks
Primary	appId	String	Unique identification strings for MHAs
	title	String	Name of MHAs
	developer	String	Developer name
	url	String	Application store URL of MHA
	description	String	Descriptions in the main MHA pages
	descriptionHTML	String	HTML format of descriptions
	summary	String	Short MHA preview page descriptions
	summaryHTML	String	HTML format of MHA summaries
	score	Integer	Average user rating
	minInstalls	Categorical	Minimum install counts of MHAs
	adSupported	Boolean{0,1}	Does the MHA support advertisement?
	containsAds	Boolean{0,1}	Does the MHA contain advertisement?
	developerEmail	Boolean{0,1}	Does the MHA have contact email?
	developerWebsite	Boolean{0,1}	Does the MHA have developer website?
	developerAddress	Boolean{0,1}	Does the MHA have developer address?
	privacyPolicy	Boolean{0,1}	Does the MHA have privacy policy page?
	icon*	String	MHA icon URL
	headerImage*	String	MHA header image URL
	screenshots*	String	MHA screenshots URLs
	video*	String	MHA video URL
	videoImage*	String	MHA video thumbnail URL
	MARS	Float	MARS quality scores of MHAs
Secondary	pred_prob_tailoring [1-15]	Float	Tailoring predicted probabilities for all the trained 15 folds
	pred_prob_selfmonitoring [1-17]	Float	Self-monitoring predicted probabilities for all the 17 trained folds
	pred_prob_praise [1-16]	Float	Praise predicted probabilities for all the trained 16 folds
	pred_prob_reminders [1-12]	Float	Reminders predicted probabilities for all the trained 12 folds

* These features are mentioned in the Future Studies section.

Table 5: Example user reviews with their textual entailment classifications

User Review	Tailoring	Self-monitoring	Praise	Reminders
This app keeps track of my blood sugar and asks, enables me to show my doctor my patterns, enables proper medication dispersion, jot notes when needed, constant reminders at specific times, keeps track of my weight, blood pressure, appointments etc...a must-have. Very much appreciative	Entails	Entails	Contradicts	Entails
App is definitely helping me track my progress, and its diabetes specific attention is much appreciated! Helpful developers when I had a problem, too. Would like to see better integration with fitness trackers like my Jawbone, and meal suggestions rather than just general advice on what to/not to eat - that would get them the last star!	Entails	Entails	Contradicts	Contradicts
Easy to keep up with has reminders easier for older people to use	Entails	Contradicts	Contradicts	Entails
I like all the features except the reminder feature	Contradicts	Contradicts	Contradicts	Entails
Great reminders, really easy to use.	Entails	Contradicts	Contradicts	Entails
I like that I can add notes and set reminders as well as track my weightloss	Entails	Entails	Contradicts	Entails
i love that everyone shares the good, bad and show support for one another. Awesome app..	Contradicts	Contradicts	Entails	Contradicts
Good glucose monitoring app clean and direct. Easy to edit.	Entails	Entails	Neutral	Neutral
Keep it simple and good. That's the secret of a good app. The missing star is for tye reminders that cannot be edited and no sync with the account on the web.	Neutral	Neutral	Neutral	Entails
I use my phone and tablet to keep track... Please sync the app! I love it and don't want to have to find another one!	Contradicts	Entails	Contradicts	Contradicts

Separate sets of user reviews were obtained for each PSD principle because user reviews are differently related to each MHA hypothesis. After textual filtering, 45, 38, 41, and 35 number of review paragraphs are created in consecutive order for *Tailoring*, *Self-monitoring*, *Praise*, and *Reminders* principles. Additionally (as in the same order), positive label counts are 15, 17, 16, and 12 for each of the mentioned PSD principles, respectively. After filtering the RD, prediction models are trained and fine-tuned. Optuna model's hyperparameter selections and their search ranges are shared in Appendix A. The finetuned model predictions in each test fold are recorded, and a single test score is calculated for each metric. Table 6 shows the results of the baselines, LightGBM, SVM, and ElasticNet classifiers.

Table 6: Binary classification results for Tailoring, Self-monitoring, Praise and Reminders principles on test folds of nested cross-validation.

Principle	Model	Accuracy	Precision	Recall	F1	MCC
Tailoring	**B1**	0.66	0.50	0.47	0.48	0.23
	B2	0.66	NA	0	NA	0
	LGB-Rev	**0.89**	**0.86**	**0.80**	**0.83**	**0.75**
	SVC-Rev	0.87	0.80	0.80	0.80	0.70
	Elast-Rev	0.87	0.85	0.73	0.79	0.69
Self-monitoring	**B1**	0.48	0.33	0.28	0.30	0.11
	B2	0.55	NA	0	NA	0
	LGB-Rev	**0.92**	**0.94**	**0.88**	**0.91**	**0.84**
	SVC-Rev	0.84	0.82	0.82	0.82	0.68
	Elast-Rev	0.87	0.80	0.94	0.87	0.75
Praise	**B1**	0.46	0.25	0.23	0.24	0.18
	B2	0.61	NA	0	NA	0
	LGB-Rev	**0.78**	**0.77**	**0.63**	**0.69**	**0.53**
	SVC-Rev	0.71	0.64	0.56	0.60	0.37
	Elast-Rev	0.68	0.59	0.63	0.61	0.34
Reminders	**B1**	0.50	0.20	0.21	0.21	0.16
	B2	0.66	NA	0	NA	0
	LGB-Rev	**0.83**	**0.95**	**0.50**	**0.67**	**0.63**
	SVC-Rev	0.74	0.67	0.50	0.57	0.40
	Elast-Rev	0.69	0.53	0.75	0.62	0.38

4.2.2 PSD Principles Prediction From Application Descriptions

Since each MHA has only one description, DD brings a smaller corpus than RD. That is why, as a qualitative check, I investigated and annotated manually all the application descriptions to determine whether these descriptions were informative. The annotations were conducted without considering the labels given by the experts after their experiments with MHAs. So that PSD principle annotation was achieved by only using the application descriptions. MHAs in DD are in the context of self-management diabetes, and annotation criteria are decided as shared in [52].

Tailoring: Adapting information for type-1 and type-2 diabetes

Self-monitoring: Feature of tracking patient health statuses such as glucose levels

Reminders: Notification sent by MHAs

Praise: Motivational compliments about user health status

After the descriptions are compared with the above criteria, certain examples show that the expert decisions are the opposite of the annotations that are derived by only looking at application descriptions. For example, in the MHA named "Diabetes World", only type-2 diabetes is mentioned as the objective disease of the MHA without mentioning type-1 diabetes. However, experts labeled the MHA as *Tailoring* principle exists. Similarly, although there is no evidence for *Reminders* and *Praise* principles, experts evaluated "Glucose Buddy" MHA as containing both principles. On the other hand, *Self-monitoring* principle was decided as "not employed" for the MHA named as "Diabetes Symptoms Causes" by the experts, although it contains the "self-monitoring" word in the description.

Further, these annotations are compared with the expert-decided PSD principle labels (true labels) overall to calculate the performance scores shared in Table 7. MCC scores (primary performance metric) are found to be very low for all the PSD principles. The performance metrics indicate that using application descriptions as a source for PSD principles' classification is noninformative.

Table 7: Annotation performance scores according to expert-evaluated true labels

	Accuracy	**Precision**	**Recall**	**F1**	**MCC**
Tailoring	0.65	0.42	0.50	0.46	0.21
Self-monitoring	0.69	0.50	0.65	0.57	0.34
Praise	0.67	0.43	0.41	0.42	0.19
Reminders	0.79	0.40	0.38	0.39	0.31

Since there is one application description for each MHA, the textual entailment step is not used for data filtering. Instead, the step is implemented to check whether its results would support the manual annotations. 3, 14, 12, and 5 mobile application descriptions (out of 75) are found to be related to the hypotheses of the *Tailoring*, *Self-monitoring*, *Praise*, and *Reminders* principles, respectively. These ratios support the qualitative findings about mobile application descriptions. Both experiments show that application descriptions lack PSD principles-related information to conduct further NLP-based classification experiments.

4.2.3 MARS Quality Score Regression

The textual entailment step (See step 2 in Figure 1) is also implemented for the MD to filter user reviews. To get the PSD predicted probabilities, the remaining user reviews are fed into the already trained PSD classification models (See step 3 in Figure 1). The predicted probabilities (as many as the fold counts) for each PSD principle are added into the MD as secondary features, which are shown in Table 4. All folds' predicted probabilities and their mean values are shown in Figure 3. The mean values of the predicted probabilities in all PSD principles appear to represent all fold values. For instance, the predicted probabilities of *Reminders* principle on different folds range between 0.6 and 0.9, and the distribution of the mean calculation of these values (at the bottom) similarly behaves. Moreover, the predicted probability distribution of *Self-monitoring* principle has bimodal distribution in different folds and hence the means of them. Thus, the mean values of predicted probabilities for each PSD principle were used in the model predictions.

In addition to the mean values of secondary features, the following MD features (See Table 4) are also used in LightGBM and CatBoost regressor models: "score", "minInstalls", "adSupported", "containAds", "developerEmail", "developerWebsite", "developerAddress", "privacyPolicy". Both regression models are fine-tuned with the Optuna framework, and the selected hyperparameters are shared in Appendix A. The baseline model performances are calculated with the mean (B3) and median (B4) values of the train set's MARS quality scores since no unsupervised regression model is found in the literature to compare with the proposed model performances. The results of the regression models in the test folds are shown in Table 8.

Table 8: MARS regression models results in test folds of nested cross-validation

Model	MAE	RMSE	Relative RMSE
B3	0.33	0.42	11.2%
B4	0.33	0.43	11.3%
LGBM-Reg (All features)	**0.27**	**0.37**	**9.7%**
LGBM-Reg (8 features)	0.29	0.38	9.8%
CatBoost-Reg (All features)	**0.29**	**0.38**	**10.1%**
CatBoost-Reg (8 features)	0.31	0.41	10.4%

To compare the LightGBM and CatBoost model performances with the B3 model (better-performing baseline), Wilcoxon-Bonferroni post hoc tests are implemented. According to comparisons, both LightGBM (RMSE = 0.37, MAE = 0.27) and CatBoost (RMSE = 0.38, MAE = 0.29) performed significantly better than the baseline model (RMSE = 0.42, MAE = 0.33), p(LightGBM - Baseline) = 0.0001, p(CatBoost - Baseline) = 0.0001.

4.2.4 Features Importance of MARS Quality Score Prediction

Overall feature importance values are calculated for both *LGBM-Reg* and *CatBoost-Reg* models over the same setup that their prediction performances calculated. Feature importance results according to information gain on their decision tree splits are shared in Table 9. Both regression model results show that there is a significant difference between the last three features and the remaining ones in terms of information gain. Moreover, SHAP (Shapley Additive Explanations) [92] values are also calculated to infer the feature importance. SHAP values are based on the feature contribution directly to the prediction model instead of a decrease in the impurity. In this way, SHAP values can provide more global values for the feature importance [92]. The obtained SHAP value summary plots for the trained regression models are shown in Figure 4. In the figure, although each feature has different distributions in terms of their effects on the outputs, their overall contributions are shown in decreasing order considering the vertical axis. “minInstalls” feature is on the top and it has the largest spread in its distribution, indicating its impact on the prediction outputs.

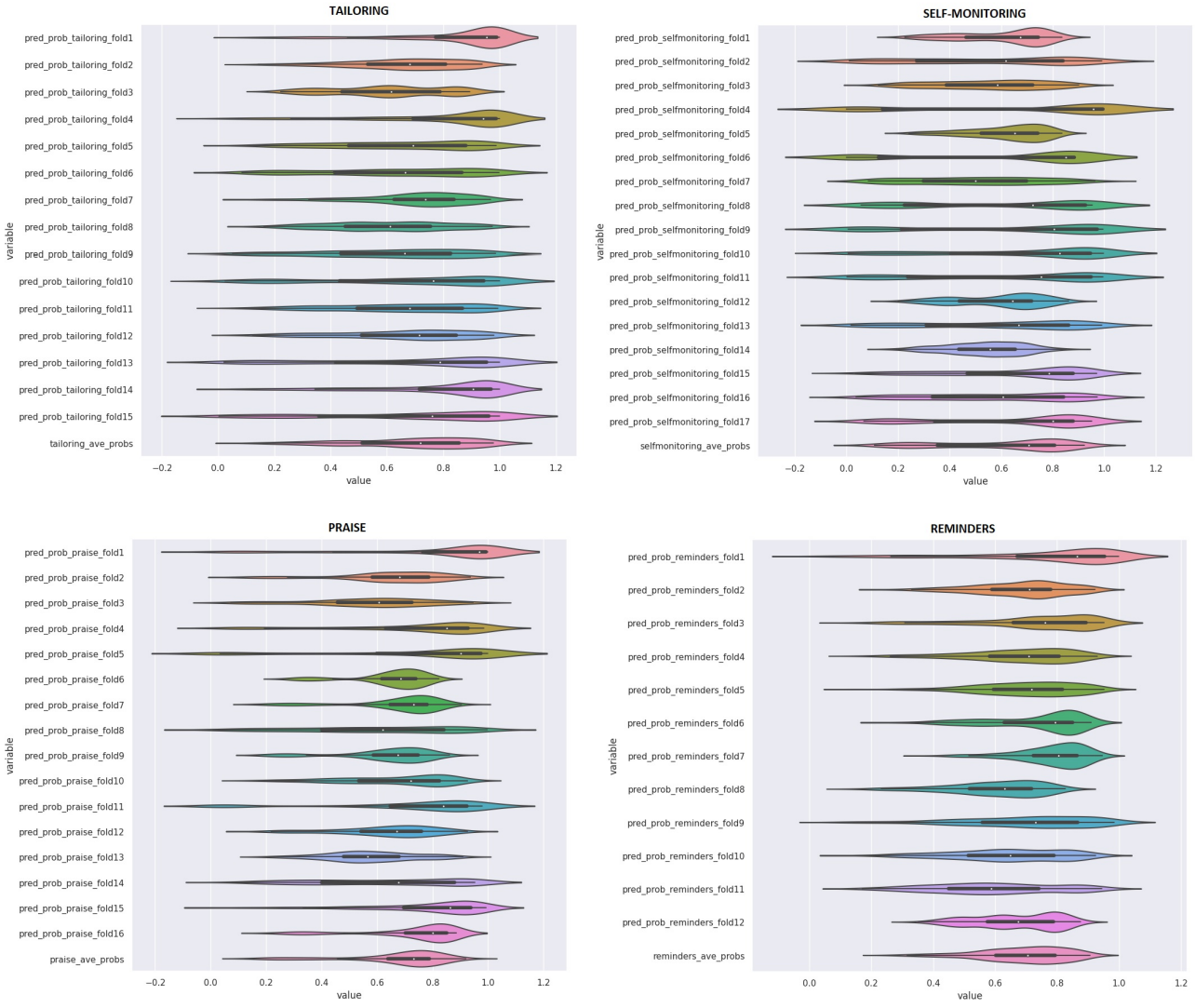

Figure 3: Violin plots of PSD predicted classification probabilities. The horizontal axis represents the distributions of predicted classification probabilities, while the vertical axis represents the fold numbers.

Table 9: Features importance obtained with information gain of *LightGBM-Reg* and *CatBoost-Reg* regression models are shown in decreasing order.

Feature	LGBM-Reg	CatBoost-Reg
minInstalls	0.123	0.266
praise_ave_probs	0.166	0.125
reminders_ave_probs	0.173	0.106
selfmonitoring_ave_probs	0.187	0.096
tailoring_ave_probs	0.129	0.111
score	0.128	0.127
containAds	0.034	0.065
adSupported	0.033	0.063
developerAddress	0.023	0.029
developerWebsite	0.003	0.005
privacyPolicy	0.001	0.007

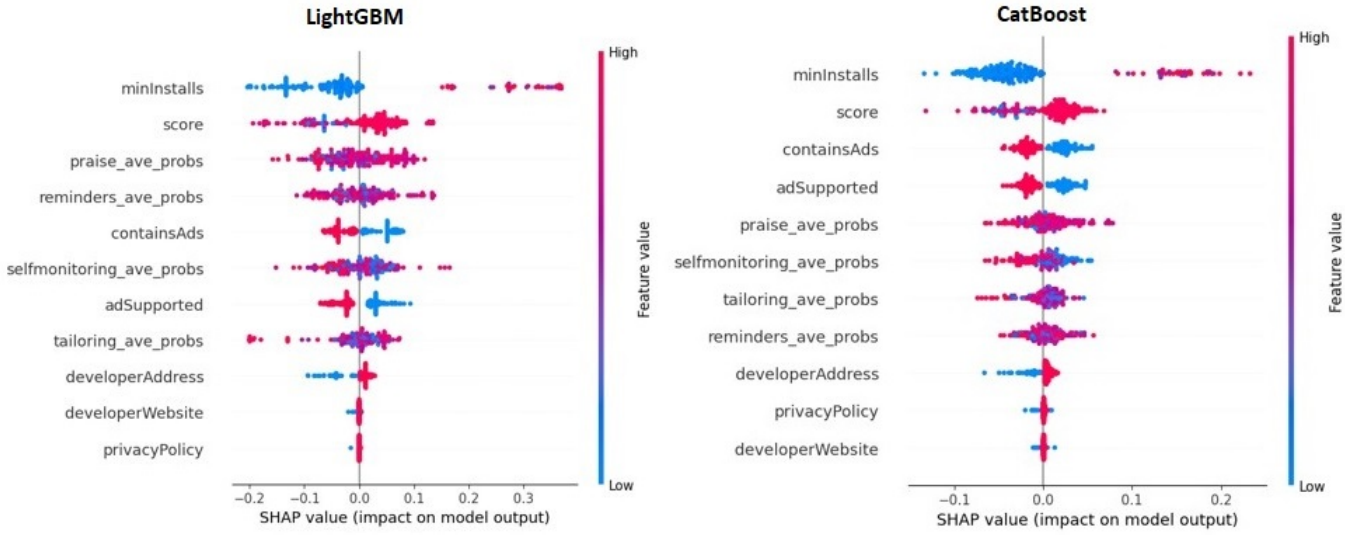

Figure 4: SHAP value summary plots obtained with LightGBM-Reg and CatBoost-Reg regression models. The vertical axis shows the feature contributions on the predictions in decreasing order. The horizontal axis shows the distribution of their impact values on the model outputs.

Considering both the calculated information gain and SHAP values, a significant decrease in the features importance is realized for the features: "developerAddress", "developerWebsite" and "privacyPolicy". Thus, LightGBM and CatBoost regression models are also experimented without these features. However, as can be seen in Table 8, the model test performances decreased, although these features have less importance compared to the other features. Moreover, the predicted PSD classification probabilities demonstrate high feature importance according to both the calculated information gain and SHAP values.

4.3 Ablation Studies

4.3.1 NLI Model Study

The ablation study is conducted to measure the effect of the selected Natural Language Inference (NLI) model on the test classification performances. XLNET (NL1) [62] and RoBERTa (NL2) [60] models are compared for the *Tailoring* principle. These models are selected because of their state-of-the-art performances in the Adversarial NLI (ANLI) benchmark [81]. The test scores for both models are calculated using the LightGBM classifier, while the other settings are kept the same. As seen in Table 10, considering the MCC scores, the XLNET model achieves a better result with its high adaptability on unseen datasets.

Table 10: Test performances of XLNET (NL1) and RoBERTa (NL2) models

Model	Accuracy	Precision	Recall	F1	MCC
NL1	**0.89**	**0.81**	**0.87**	**0.84**	**0.75**
NL2	0.82	0.90	0.60	0.72	0.62

4.3.2 NLI Hypothesis Study

In this ablation study, the effect of the used hypotheses for the textual entailment step (See step 2 in Figure 1) is examined. Each of the compared hypotheses in this section is generated regarding the given descriptions and examples about PSD principles. They are used as hypotheses in the NLI model of XLNET [62] to filter user reviews for the *Reminders* and *Tailoring* principles. The filtered user reviews are classified using LightGBM classifiers, which are identified as the best-performing model in Section 4.2.1, and the other setups for the experiments are kept the same. The first and second hypotheses are created to compare the effect of using two synonyms words: "notification" and "messages".

1. The text is about sending notifications to users.
2. The text is about sending messages to users.

Using a direct mobile applications-related term (“notification”) instead of a more general noun (“reminder”) resulted in higher performance for the classification, as shown in Table 11. Further, a third hypothesis is created by extending the first one (the higher performer hypothesis) to see the effect of defining clauses in the hypothesis.

3. The text is about an application that sends notifications to users.

Although the third hypothesis is more appealing and grammatically correct, the classification test performance is significantly dropped as can be seen in Table 11. In the second hypothesis, a higher result is achieved using a shorter sentence and mobile application-related feature name.

Table 11: *Reminders* principle’s classification test performances according to the proposed hypotheses.

Hypothesis	Accuracy	Precision	Recall	F1	MCC
1	**0.83**	**0.95**	**0.50**	**0.67**	**0.63**
2	0.80	0.75	0.64	0.69	0.55
3	0.72	0.67	0.36	0.47	0.33

Moreover, *“tailored”* and *“adaptive”* words are also compared in the hypotheses of the *Tailoring* principle. Although “tailoring” is the PSD principle’s name, its first and second meanings are defined as “the business or work of a tailor” and “the skill or craftsmanship of a tailor” [95]. Thus, the below hypotheses are compared to see the effect of using words in their connotation meaning for the textual entailment task.

4. The text is about providing tailored information for users.
5. The text is about providing adaptive information for users.

The results in Table 12 show that using words with their connotation meaning decreases the PSD classification performance.

Table 12: *Tailoring* principle’s classification test performances according to the created hypotheses

Hypothesis	Accuracy	Precision	Recall	F1	MCC
4	0.87	0.91	0.67	0.77	0.70
5	**0.89**	**0.86**	**0.80**	**0.83**	**0.75**

CHAPTER 5

DISCUSSION

In this section, the experiments findings are further discussed, as well as possible future works are explained for each research question.

5.1 Experiments Results

5.1.1 PSD Principles Prediction

RQ1 How can we predict PSD principles of an expert-evaluated MHA based on user reviews and application descriptions?

It is hypothesized that user reviews and application descriptions include the key attributes, functionalities, and objectives of the MHA as perceived by users and developers. Therefore, conducting a content analysis on these sources can provide valuable insights into the utilized principles of PSD in MHAs. To explore this, large language models are employed for the textual entailment classifications of both user reviews and application descriptions with the definitions of PSD principles.

The textual entailment results show that user reviews can bring enough records that entail the PSD definitions. The textual entailment step is employed to filter the posted user reviews, which are then concatenated to form paragraphs defining each MHA. These paragraphs are used to generate sentence embeddings, which serve as the feature sets for binary classification models. By utilizing these feature sets, PSD principles can be predicted with higher accuracy compared to the baseline models. Therefore, user reviews can be utilized for predicting PSD principles. However, the prediction performance varies for each experimented PSD principle, as presented in Table6.

While significantly high Matthews Correlation Coefficient (MCC) scores are achieved in predicting the *Tailoring* and *Self-monitoring* principles, the prediction for the *Praise* principle is underperforming. This can be attributed to the differences in the original definitions of PSD principles. During the expert evaluations, these principles are identified based on predefined features, which are exemplified in Table 2. Upon manual investigation of positively labeled MHAs for the *Praise* principle, it is observed that users do not mention motivational feedback provided by the applications. Conversely, there is evidence supporting the presence of *Tailoring* and *Self-monitoring*

principles in user reviews. As can be seen in Table 5, user reviews such as *"... diabetes-specific attention.."* and *"...keeps track of my blood sugar..."* have directly relatable keywords ("specific" and "track") with the definitions of PSD principles. On the other hand, the low performance of the *Reminders* principle compared to the others can be explained by its having lower positive labeled MHA in the extracted datasets. As shown in Table 3, the *Reminders* principle has the lowest implementation ratio among the experimented principles. Consequently, the prediction models are trained with fewer positive labels and fewer folds in the nested stratified cross-validation ensemble method. The lower prediction performance for the *Reminders* principle further justifies excluding other PSD principles with low implementation ratios from the study. On the other hand, a low entailment ratio is observed between application descriptions and PSD principle definitions. The textual entailment results of application descriptions are cross-verified with manual annotations, which support the classification results of textual entailments, as indicated in Table 7. Therefore, the steps implemented for user reviews are not experimented with using application descriptions, as they do not provide informative insights regarding PSD principles.

5.1.2 MARS Quality Score Predictions

RQ2 How can we predict MHA quality on the scale of MARS by using the collected descriptive MHA information? Can PSD principles that are predicted by the models contribute to this prediction?

To answer this research question two-step experiment is conducted. Firstly, supervised regression models are developed using the descriptive features of mobile applications and the predicted classification probabilities obtained from PSD principle classifications. High test performances are achieved for the experimented regression models, as can be seen in Table 8. Secondly, the features importance for these models are calculated using the information gain and SHAP value methods. The information gain scores indicate that all four predicted classification probabilities of PSD principles rank within the top six in terms of importance, while according to the SHAP values, they rank within the top eight. The used decision tree methods are biased toward high cardinality categorical features (e.g., install count) in terms of their gradient boosting natures [96,97]. Considering this bias, having close feature importance to the categorical install count feature shows that predicted classification probabilities of PSD principles contribute to MARS quality predictions.

Furthermore, additional tests are conducted by predicting mobile application qualities using the same models but excluding the three least important features (developerAddress, developerWebsite, privacyPolicy). The performance of the regression models decreased with the exclusion of these features, as demonstrated in Table 8. According to the SHAP values, all the predicted classification probabilities exhibit higher feature importance compared to the removed features. The SHAP values directly represent the importance of features in relation to the prediction outputs. Thus, these experiments demonstrate that all predicted classification probabilities of PSD principles contribute to the prediction of MARS quality scores.

5.2 Future Work

The improvement of large language model performances can be achieved through finetuning or prompt tuning, which is further training of pre-trained models to respond to specific tasks [98]. However, in this study, models are utilized without these methods due to the unavailability of datasets containing mobile application-related user reviews or descriptions. For instance, manual inspection of the classification outputs in the textual entailment step (see step 2 in Figure 1) revealed significant instances of false positive classifications, particularly in the case of short user reviews. An example of such a false positive classification is the user review *"Great reminders, really easy to use."* which is erroneously classified as entailed with the *Tailoring* principle (see Table 5), although outputs for *Self-monitoring*, *Reminders* and *Praise* principles are correct. As the selected NLI model used in this study is trained on longer texts, their performance suffers when classifying short user reviews. Although there is no user data specific to medical applications, these models will be further trained in the future using publicly available user review datasets such as [99] and [100] to achieve better performance on shorter user reviews.

In this research, an analysis was conducted on user reviews in a holistic manner, without differentiating them based on user sentiments or user ratings. However, it is worth noting that user reviews expressing negative sentiments often provide more specific references to application features. Therefore, in future studies, it would be valuable to explore the relation between the information content and sentiments or user rating of user reviews, as they can offer valuable insights. The relation can be inferred from the prediction of PSD principles based on only negative user reviews.

Furthermore, with the manual inspections of application descriptions, it is realized that information about the subcategories of mobile applications can be obtained. While all the included applications in this study fall under the general "medical" or "health" categories, a significant portion of them primarily provide tips for better living rather than incorporating health management functionalities that align with PSD principles. Notably, regarding their application descriptions, it was observed that out of the 76 evaluated MHAs, 11 are specifically developed to provide dietary advice for diabetic patients. Furthermore, among these applications, 7 do not incorporate any PSD principles, as identified by both the author and expert evaluators. Therefore, future studies should aim to extract and utilize subcategories related to the nature of MHAs to gain a deeper understanding of their characteristics and potential purposes. While the extracted subcategories may contribute to health application analysis, they can be a distinguishing feature in mobile application research in general.

MARS [9] provides a questionnaire that includes 19 Likert-scaled questions used to assess various aspects of MHAs, such as engagement, functionality, aesthetics, and functionality subscores. Although the current study utilized descriptive application store features of MHAs to predict quality scores, the inclusion of these Likert-scaled questions in the development of question-answer models has potential. By utilizing the extracted datasets such as MD, RD, and DD, future studies can explore the performance of question-answer models by mining textual data and compare them with

the regression models developed in this thesis. Incorporating the questionnaire-based approach can provide a more comprehensive evaluation of MHA qualities, capturing user perspectives similar to PSD classifications.

The creation of medical application datasets is a contribution of this study. These datasets include various visual data fields such as application icons, header images, screenshots, application videos, and video images, which were collected and saved (see Table 4). However, the quality of the visuals and the performance of the OCRs used for keyword extraction were found to be insufficient for inclusion in the feature sets. Nevertheless, future research can reevaluate these visual data with the use of emerging image analysis for feature extraction purposes.

CHAPTER 6

CONCLUSION

In summary, this study attempts to propose automatic evaluation techniques for mobile health applications according to persuasive system design principles and mobile application rating scale. For this purpose, mobile health application descriptions and posted user reviews are scraped and enriched with other descriptive information about applications. Large language models are utilized with supervised machine learning models to predict the assessment of experts regarding implemented PSD principles of mobile health applications. Furthermore, output predicted classification probabilities are utilized during the regression of mobile health application qualities, and their features importance are calculated to infer the relation between PSD principles and MARS quality scores of MHAs. These steps are conducted to answer the following research questions.

RQ1 How can we predict PSD principles of an expert-evaluated MHA based on user reviews and application descriptions?

It is hypothesized that collected user reviews and application descriptions contain the MHA's main features, functionalities, and purposes from the perspectives of users and developers. Hence, user reviews and application descriptions can be relevant sources for conducting content analysis regarding employed PSD principles of MHAs. Textual entailment classifications show that user reviews are informative in terms of PSD principles. Utilizing large language models and supervised machine learning classifiers, *Tailoring*, *Self-monitoring*, *Praise*, and *Reminders* principles can be predicted with higher performance scores than baseline models. From these PSD principles, *Tailoring* and *Self-monitoring* are predicted with significantly high MCC scores, while *Praise* and *Reminders* underperformed. Different performances of PSD principles can be explained by the differences in positive label ratios of generated datasets and the lack of mentions in user reviews.

RQ2 How can we predict MHA quality on the scale of MARS by using the collected descriptive MHA information? Can PSD principles that are predicted by the models contribute to this prediction?

It is hypothesized that collected descriptive data (e.g., overall customer rating, install counts, advertisement support) and predicted classification probabilities of PSD principles could create a feature vector to predict MARS overall scores. Trained supervised regression models predict the quality scores of MHAs demonstrating high

performance. Features importance of these models according to information gain and SHAP values indicate that the predicted PSD classification probabilities significantly contribute to the prediction of overall MARS quality scores of MHAs.

The study contributes to the mobile health applications field by addressing the gap in the literature regarding the effects of PSD principles on mobile health application qualities. The findings of this study have implications not only for researchers but also for mobile health application developers. Although the study experimented on already developed and published MHAs, conclusions are also valid for user-centric new mobile health application developments. By incorporating PSD principles during application development steps, user engagement with high-quality applications can be achieved by developers. Moreover, developed applications can be monitored and further improved by the developers with the proposed evaluation techniques in this study.

6.1 Limitations and Assumptions

During the process of data collection, application descriptions and user reviews are gathered from the public application store as in their latest versions. The user reviews are filtered based on their posting dates to exclude newer reviews posted after the expert evaluations. However, the application descriptions do not include version or posting date information, and only the most recent description available from the application store is accessible. These application descriptions are collected shortly after the expert annotations are published in the study by Geirhos et al. [52]. It is assumed that there have been no significant changes to the application descriptions during this time period and experiments are conducted on the available data.

During the data evaluation, *Tailoring*, *Self-monitoring*, *Praise* and *Reminders* principles are selected to conduct experiments. The selection of these principles is based on their positively labeled MHA ratios and their correlation with the MARS quality scores. Verifiability and third-party endorsement principles, despite having promising ratios (see Table 3), are not selected since they have been reported to be less correlated with the target MARS scores [52]. Moreover, these principles do not have definitions that can be inferred from user reviews. Both verifiability (*"information sources should be shared to verify"*) and third-party endorsement (*"endorsements/certificates should be provided from respected sources"*) are related to clear MHA requirements of referring sources. These can directly search from MHA pages with simple queries instead of analysis conducted in this study.

All the code experiments are conducted using Google Colab Pro VMs that have the following system settings: *GPU: Tesla T4 (UUID: GPU-9f6e161b-db80-86e5-040c-dbf6c0b5f55f), NVIDIA-SMI 525.85.12, Driver Version: 525.85.12, CUDA Version: 12.0, Intel XI(R) CPU @ 2.30GHz, L3 cache: 45MB, 26G Available Memory.*

Although the paid version of Google Colab is used, there is limited available memory for high-speed computing. Consequently, for the sentence embedding task, the ST5-Large model is utilized instead of the larger ST5-XL and ST5-XXL models from the same family. The ST5-Large model demonstrates significant performance in the text classification tasks according to the Massive Text Embedding Benchmark (MTEB) [83]. Although ST5-Large model has lower performance than its larger versions, it is faster [82]. The fast performance is also important during the model selection since sentence embeddings are generated multiple times in different models and PSD principles experiments.

In the prediction of the employed PSD principles of MHAs, the aim is to achieve expert-level predictions, and the performances are evaluated based on expert annotations for the MHAs. However, in this thesis, the automatic predictions are designed in a more generalized manner compared to the study conducted by Geirhos et al. [52]. In their study, experts were provided with diabetes-specific features to make determinations regarding the employed PSD principles. For instance, the *Tailoring* principle was considered employed if the MHA provided adaptive information for both type-1 and type-2 diabetes. However, in this thesis, a more generalized detection method is implemented, independent of specific disease details, in order to propose a more generalized assessment of MHAs and contribute to the broader field of medical health applications. Therefore, in the textual entailment step, hypotheses are formulated without disease-specific definitions, assuming that they can perform similarly across specific disease applications.

REFERENCES

[1] S. Kumar, W. J. Nilsen, A. Abernethy, A. Atienza, K. Patrick, M. Pavel, W. T. Riley, A. Shar, B. Spring, D. Spruijt-Metz, *et al.*, "Mobile health technology evaluation: the mhealth evidence workshop," *American journal of preventive medicine*, vol. 45, no. 2, pp. 228–236, 2013.

[2] J. Linardon, P. Cuijpers, P. Carlbring, M. Messer, and M. Fuller-Tyszkiewicz, "The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials," *World Psychiatry*, vol. 18, no. 3, pp. 325–336, 2019.

[3] S. Cerea, G. Doron, T. Manoli, F. Patania, G. Bottesi, and M. Ghisi, "Cognitive training via a mobile application to reduce some forms of body dissatisfaction in young females at high-risk for body image disorders: A randomized controlled trial," *Body image*, vol. 42, pp. 297–306, 2022.

[4] O. Byambasuren, S. Sanders, E. Beller, and P. Glasziou, "Prescribable mhealth apps identified from an overview of systematic reviews," *NPJ digital medicine*, vol. 1, no. 1, p. 12, 2018.

[5] Statista, "Number of mhealth apps available in the apple app store from 1^{st} quarter 2015 to 3^{rd} quarter 2022." `https://www.statista.com/statistics/779910/health-apps-available-ios-worldwide/`, 2022. (accessed on May 5, 2023.

[6] Statista, "Number of available health apps in google play worldwide from 3^{rd} quarter 2014 to 3^{rd} quarter 2022." `https://www.statista.com/statistics/779919/health-apps-available-google-play-worldwide/`, 2022. (accessed on May 5, 2023.

[7] S. Akbar, E. Coiera, and F. Magrabi, "Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review," *Journal of the American Medical Informatics Association*, vol. 27, no. 2, pp. 330–340, 2020.

[8] H. Oinas-Kukkonen and M. Harjumaa, "Persuasive systems design: Key issues, process model, and system features," *Communications of the association for Information Systems*, vol. 24, no. 1, p. 28, 2009.

[9] S. R. Stoyanov, L. Hides, D. J. Kavanagh, O. Zelenko, D. Tjondronegoro, and M. Mani, "Mobile app rating scale: a new tool for assessing the quality of health mobile apps," *JMIR mHealth and uHealth*, vol. 3, no. 1, p. e3422, 2015.

[10] F. Palomba, P. Salza, A. Ciurumelea, S. Panichella, H. Gall, F. Ferrucci, and A. De Lucia, "Recommending and localizing change requests for mobile apps based on user reviews," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, pp. 106–117, IEEE, 2017.

[11] A. Ciurumelea, A. Schaufelbühl, S. Panichella, and H. C. Gall, "Analyzing reviews and code of mobile apps for better release planning," in *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pp. 91–102, IEEE, 2017.

[12] N. Chen, J. Lin, S. C. Hoi, X. Xiao, and B. Zhang, "Ar-miner: mining informative reviews for developers from mobile app marketplace," in *Proceedings of the 36th international conference on software engineering*, pp. 767–778, 2014.

[13] L. Villarroel, G. Bavota, B. Russo, R. Oliveto, and M. Di Penta, "Release planning of mobile apps based on user reviews," in *Proceedings of the 38th International Conference on Software Engineering*, pp. 14–24, 2016.

[14] S. Scalabrino, G. Bavota, B. Russo, M. Di Penta, and R. Oliveto, "Listening to the crowd for the release planning of mobile apps," *IEEE Transactions on Software Engineering*, vol. 45, no. 1, pp. 68–86, 2017.

[15] A. Di Sorbo, S. Panichella, C. V. Alexandru, C. A. Visaggio, and G. Canfora, "Surf: summarizer of user reviews feedback," in *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, pp. 55–58, IEEE, 2017.

[16] E. Noei, F. Zhang, and Y. Zou, "Too many user-reviews! what should app developers look at first?," *IEEE Transactions on Software Engineering*, vol. 47, no. 2, pp. 367–378, 2019.

[17] C. Stanik, M. Haering, and W. Maalej, "Classifying multilingual user feedback using traditional machine learning and deep learning," in *2019 IEEE 27th international requirements engineering conference workshops (REW)*, pp. 220–226, IEEE, 2019.

[18] N. Al Kilani, R. Tailakh, and A. Hanani, "Automatic classification of apps reviews for requirement engineering: Exploring the customers need from healthcare applications," in *2019 sixth international conference on social networks analysis, management and security (SNAMS)*, pp. 541–548, IEEE, 2019.

[19] H. Wu, W. Deng, X. Niu, and C. Nie, "Identifying key features from app user reviews," in *2021 IEEE/ACM 43th International Conference on Software Engineering (ICSE)*, pp. 922–932, IEEE, 2021.

[20] T. Johann, C. Stanik, W. Maalej, *et al.*, "Safe: A simple approach for feature extraction from app descriptions and app reviews," in *2017 IEEE 25th international requirements engineering conference (RE)*, pp. 21–30, IEEE, 2017.

[21] N. Chen, S. C. Hoi, S. Li, and X. Xiao, "Simapp: A framework for detecting similar mobile applications by online kernel learning," in *Proceedings of the*

eighth ACM international conference on web search and data mining, pp. 305–314, 2015.

[22] B. Hu, B. Liu, N. Z. Gong, D. Kong, and H. Jin, "Protecting your children from inappropriate content in mobile apps: An automatic maturity rating framework," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1111–1120, 2015.

[23] M. Liu, H. Wang, Y. Guo, and J. Hong, "Identifying and analyzing the privacy of apps for kids," in *Proceedings of the 17th international workshop on mobile computing systems and applications*, pp. 105–110, 2016.

[24] "Google play developer distribution agreement." `https://play.google.com/about/developer-distribution-agreement.html`.

[25] "App store review guidelines." `https://developer.apple.com/app-store/review/guidelines/`.

[26] M. Hood, R. Wilson, J. Corsica, L. Bradley, D. Chirinos, and A. Vivo, "What do we know about mobile applications for diabetes self-management? a review of reviews," *Journal of behavioral medicine*, vol. 39, pp. 981–994, 2016.

[27] D. S. Eng and J. M. Lee, "The promise and peril of mobile health applications for diabetes and endocrinology," *Pediatric diabetes*, vol. 14, no. 4, pp. 231–238, 2013.

[28] P. Llorens-Vernet, J. Miró, *et al.*, "The mobile app development and assessment guide (mag): Delphi-based validity study," *JMIR mHealth and uHealth*, vol. 8, no. 7, p. e17760, 2020.

[29] M. Aitken and C. Gauntlett, "Patient apps for improved healthcare: from novelty to mainstream," *Parsippany, NJ: IMS Institute for Healthcare Informatics*, 2013.

[30] A. Domnich, L. Arata, D. Amicizia, A. Signori, B. Patrick, S. Stoyanov, L. Hides, R. Gasparini, and D. Panatto, "Development and validation of the Italian version of the mobile application rating scale and its generalisability to apps targeting primary prevention," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, pp. 1–10, 2016.

[31] E.-M. Messner, Y. Terhorst, A. Barke, H. Baumeister, S. Stoyanov, L. Hides, D. Kavanagh, R. Pryss, L. Sander, T. Probst, *et al.*, "The german version of the mobile app rating scale (mars-g): development and validation study," *JMIR mHealth and uHealth*, vol. 8, no. 3, p. e14479, 2020.

[32] T. L. Lewis and J. C. Wyatt, "mhealth and mobile medical apps: a framework to assess risk and promote safer use," *Journal of medical Internet research*, vol. 16, no. 9, p. e3133, 2014.

[33] M. Bardus, S. B. van Beurden, J. R. Smith, and C. Abraham, "A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 13, no. 1, pp. 1–9, 2016.

[34] S. Jovičić, J. Siodmiak, I. D. Watson, E. F. of Clinical Chemistry, and L. M. W. G. on Patient Focused Laboratory Medicine, "Quality evaluation of smartphone applications for laboratory medicine," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 57, no. 3, pp. 388–397, 2019.

[35] S. R. N. Kalhori, M. Hemmat, T. Noori, S. Heydarian, and M. R. Katigari, "Quality evaluation of english mobile applications for gestational diabetes: app review using mobile application rating scale (mars)," *Current Diabetes Reviews*, vol. 17, no. 2, pp. 161–168, 2021.

[36] G. Gross, C. Lull, J. von Ahnen, V. Olsavszky, J. Knitza, A. Schmieder, and J. Leipe, "German mobile apps for patients with psoriatic arthritis: Systematic app search and content analysis," *Health Policy and Technology*, vol. 11, no. 4, p. 100697, 2022.

[37] B. Y. Kim, A. Sharafoddini, N. Tran, E. Y. Wen, and J. Lee, "Consumer mobile apps for potential drug-drug interaction check: systematic review and content analysis using the mobile app rating scale (mars)," *JMIR mHealth and uHealth*, vol. 6, no. 3, p. e8613, 2018.

[38] S. N. Selvaraj and A. Sriram, "The quality of indian obesity-related mhealth apps: Precede-proceed model–based content analysis," *JMIR mHealth and uHealth*, vol. 10, no. 5, p. e15719, 2022.

[39] B. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[40] B. J. Fogg, "A behavior model for persuasive design," in *Proceedings of the 4^{th} international Conference on Persuasive Technology*, pp. 1–7, 2009.

[41] R. B. Cialdini, "Harnessing the science of persuasion," *Harvard Business Review*, vol. 79, no. 9, pp. 72–79, 2001.

[42] C. Abraham and S. Michie, "A taxonomy of behavior change techniques used in interventions.," *Health psychology*, vol. 27, no. 3, p. 379, 2008.

[43] J. Geuens, T. W. Swinnen, R. Westhovens, K. De Vlam, L. Geurts, V. V. Abeele, *et al.*, "A review of persuasive principles in mobile apps for chronic arthritis patients: opportunities for improvement," *JMIR mHealth and uHealth*, vol. 4, no. 4, p. e6286, 2016.

[44] R. A. Asbjørnsen, J. Wentzel, M. L. Smedsrød, J. Hjelmesæth, M. M. Clark, L. Solberg Nes, and J. E. Van Gemert-Pijnen, "Identifying persuasive design principles and behavior change techniques supporting end user values and

needs in ehealth interventions for long-term weight loss maintenance: qualitative study," *Journal of medical Internet research*, vol. 22, no. 11, p. e22598, 2020.

[45] R. A. Asbjørnsen, J. Hjelmesæth, M. L. Smedsrød, J. Wentzel, M. Ollivier, M. M. Clark, J. E. van Gemert-Pijnen, and L. Solberg Nes, "Combining persuasive system design principles and behavior change techniques in digital interventions supporting long-term weight loss maintenance: design and development of echange," *JMIR human factors*, vol. 9, no. 2, p. e37372, 2022.

[46] E. Kaasinen, E. Mattila, H. Lammi, T. Kivinen, and P. Välkkynen, "Technology acceptance model for mobile services as a design framework," in *Human-Computer interaction and innovation in handheld, mobile and wearable technologies*, pp. 80–107, IGI Global, 2011.

[47] O. Oyebode, C. Ndulue, M. Alhasani, and R. Orji, "Persuasive mobile apps for health and wellness: a comparative systematic review," in *Persuasive Technology. Designing for Future Change: 15th International Conference on Persuasive Technology, PERSUASIVE 2020, Aalborg, Denmark, April 20–23, 2020, Proceedings 15*, pp. 163–181, Springer, 2020.

[48] F. Alqahtani, G. Al Khalifah, O. Oyebode, and R. Orji, "Apps for mental health: an evaluation of behavior change strategies and recommendations for future development," *Frontiers in Artificial Intelligence*, vol. 2, p. 30, 2019.

[49] S. Langrial, T. Lehto, H. Oinas-Kukkonen, M. Harjumaa, and P. Karppinen, "Native mobile applications for personal well-being: a persuasive systems design evaluation," 2012.

[50] M. A. Al-Ramahi, J. Liu, and O. F. El-Gayar, "Discovering design principles for health behavioral change support systems: A text mining approach," *ACM Transactions on Management Information Systems (TMIS)*, vol. 8, no. 2-3, pp. 1–24, 2017.

[51] O. Oyebode and R. Orji, "Deconstructing persuasive strategies in mental health apps based on user reviews using natural language processing.," in *BCSS@ PERSUASIVE*, 2020.

[52] A. Geirhos, M. Stephan, M. Wehrle, C. Mack, E.-M. Messner, A. Schmitt, H. Baumeister, Y. Terhorst, and L. Sander, "Standardized evaluation of the quality and persuasiveness of mobile health applications for diabetes management," *Scientific reports*, vol. 12, no. 1, pp. 1–10, 2022.

[53] E. D. Liddy, "Natural language processing," *M.A. Drake (Ed.), Encyclopedia of library and information science (2nd ed.), Marcel Decker, Inc.*, 2001.

[54] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia tools and applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[55] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. MIT press, 1999.

[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[58] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[59] 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*, 2019.

[60] 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.

[61] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[62] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[63] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[64] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[65] T. A. Koleck, C. Dreisbach, P. E. Bourne, and S. Bakken, "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review," *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 364–379, 2019.

[66] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet, *et al.*, "Machine learning and natural language processing in mental health: systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, p. e15708, 2021.

[67] E. Sezgin, S.-A. Hussain, S. Rust, and Y. Huang, "Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: Feasibility study with real-world data," *JMIR Formative Research*, vol. 7, p. e43014, 2023.

[68] S. Meystre and P. J. Haug, "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation," *Journal of biomedical informatics*, vol. 39, no. 6, pp. 589–599, 2006.

[69] R. Crutzen, G.-J. Y. Peters, S. D. Portugal, E. M. Fisser, and J. J. Grolleman, "An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study," *Journal of Adolescent Health*, vol. 48, no. 5, pp. 514–519, 2011.

[70] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. Lau, *et al.*, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[71] R. Garg, E. Oh, A. Naidech, K. Kording, and S. Prabhakaran, "Automating ischemic stroke subtype classification using machine learning and natural language processing," *Journal of Stroke and Cerebrovascular Diseases*, vol. 28, no. 7, pp. 2045–2051, 2019.

[72] J. H. Garvin, Y. Kim, G. T. Gobbel, M. E. Matheny, A. Redd, B. E. Bray, P. Heidenreich, D. Bolton, J. Heavirland, N. Kelly, *et al.*, "Automating quality measures for heart failure using natural language processing: a descriptive study in the department of veterans affairs," *JMIR medical informatics*, vol. 6, no. 1, p. e9150, 2018.

[73] E. Pons, L. M. Braun, M. M. Hunink, and J. A. Kors, "Natural language processing in radiology: a systematic review," *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[74] S. Dutta, W. J. Long, D. F. Brown, and A. T. Reisner, "Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings," *Annals of emergency medicine*, vol. 62, no. 2, pp. 162–169, 2013.

[75] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "Brat: a web-based tool for nlp-assisted text annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107, 2012.

[76] M. Jo, "Google play scraper v1.2.3." `https://github.com/JoMingyu/google-play-scraper`. (accessed on Apr. 21, 2023).

[77] S. Han, "googletrans v3.0.0." `https://github.com/ssut/py-googletrans`. (accessed on Apr. 21, 2023).

[78] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[79] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122, Association for Computational Linguistics, 2018.

[80] Y. Nie, H. Chen, and M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[81] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial nli: A new benchmark for natural language understanding," *arXiv preprint arXiv:1910.14599*, 2019.

[82] J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.

[83] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022.

[84] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[85] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[86] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[87] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "A comparison of decision tree ensemble creation techniques," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 1, pp. 173–180, 2006.

[88] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, pp. 261–283, 2013.

[89] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25^{th} ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

[90] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, pp. 1–13, 2020.

[91] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[92] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[93] Apple Inc., "App store improvements." https://developer.apple.com/support/app-store-improvements. (accessed on Apr. 25, 2023).

[94] M. Stach, R. Kraft, T. Probst, E.-M. Messner, Y. Terhorst, H. Baumeister, M. Schickler, M. Reichert, L. B. Sander, and R. Pryss, "Mobile health app database-a repository for quality ratings of mhealth apps," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 427–432, IEEE, 2020.

[95] Collins, "tailoring." https://www.collinsdictionary.com/us/dictionary/english/tailoring. (accessed on Apr. 30, 2023).

[96] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–21, 2007.

[97] A. I. Adler and A. Painsky, "Feature importance in gradient boosting trees with cross-validation feature selection," *Entropy*, vol. 24, no. 5, p. 687, 2022.

[98] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H.-T. Zheng, and M. Sun, "Openprompt: An open-source framework for prompt-learning," *arXiv preprint arXiv:2111.01998*, 2021.

[99] M. M. Marchetti, "Play store sentiment analysis of user reviews notebook." https://www.kaggle.com/code/mmmarchetti/play-store-sentiment-analysis-of-user-reviews/notebook, 2019. (accessed on May. 9, 2023).

[100] P. Rathi, "Google play store reviews." https://www.kaggle.com/datasets/prakharrathi25/google-play-store-reviews, 2021. (accessed on May. 9, 2023).

APPENDIX A

OPTUNA HYPERPARAMETER TUNING

Table 13: Hyperparameter tunings search ranges and the selected values by Optuna for LGBM-Rev models.

Parameters	Search Range	Selected Values			
		Tailoring	Self-monitoring	Praise	Reminders
learning_rate	[0.001, 0.1]	0.05	0.04	0.04	0.03
lambda_l1	$[10^{-8}, 10]$	0.01	0.02	0.18	0.15
lambda_l2	$[10^{-8}, 10]$	0.09	0.03	0.01	0.48
num_leaves	[2, 512]	254	299	168	129
feature_fraction	[0.5, 1.0]	0.72	0.91	0.86	0.71
bagging_fraction	[0.5, 1.0]	0.83	0.84	0.87	0.71
bagging_freq	[0, 15]	8	10	7	7
min_child_samples	[1, 32]	9	9	5	6

Table 14: Hyperparameter tuning search ranges and the selected values by Optuna for LGBM-Reg models.

Parameters	Search Range	Selected Values	
		All Features	8 features
learning_rate	[0.001, 0.1]	0.03	0.03
lambda_l1	$[10^{-8}, 10]$	0.34	0.53
lambda_l2	$[10^{-8}, 10]$	1.85	0.09
num_leaves	[2, 512]	270	300
feature_fraction	[0.5, 1.0]	0.68	0.78
bagging_fraction	[0.5, 1.0]	0.66	0.61
bagging_freq	[0, 15]	7	3
min_child_samples	[1, 32]	9	12

Table 15: Hyperparameter tuning search ranges and the selected values by Optuna for CatBoost-Reg models.

Parameters	Search Range	Selected Values	
		All Features	8 features
learning_rate	[0.001, 0.1]	0.05	0.07
depth	[4, 32]	13	12
l2_leaf_reg	[1, 8]	2	2.5
min_child_samples	[1, 32]	8	8