

Titre: Understanding User Perception of Mobile App Quality: A Large-Scale Multi-Dimension Study of the Reviews and Ratings of Android Sports Apps
Title:

Auteur: Bhagya Chembakottu
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Chembakottu, B. (2023). Understanding User Perception of Mobile App Quality: A Large-Scale Multi-Dimension Study of the Reviews and Ratings of Android Sports Apps [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie.
Citation: <https://publications.polymtl.ca/53358/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/53358/>
PolyPublie URL:

Directeurs de recherche: Heng Li, & Foutse Khomh
Advisors:

Programme: Génie informatique
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Understanding User Perception of Mobile App Quality:
A Large-scale Multi-dimension Study of the Reviews and Ratings of Android Sports Apps**

BHAGYA CHEMBAKOTTU

Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Avril 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Understanding User Perception of Mobile App Quality:
A Large-scale Multi-dimension Study of the Reviews and Ratings of Android Sports Apps**

présenté par **Bhagya CHEMBAKOTTU**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Michel DESMARAIS, président

Heng LI, membre et directeur de recherche

Foutse KHOMH, membre et codirecteur de recherche

Mohammad HAMDAQA, membre

DEDICATION

*To my Amma, Achan and Ammu, who supported me in pursuing my dreams,
my friends in India who educated me with opportunities and encouraged me to get there,
Appu and Apollo, who became family here and supported and believed in me,
and friends and lab mates,
who believed in me and supported me for my graduate study journey . . .*

ACKNOWLEDGEMENTS

First and foremost, I express my deep appreciation to my research director, Heng Li, for providing me with the opportunity to pursue my graduate studies under his supervision. His unwavering belief in me as his student and his invaluable guidance throughout my research work has been instrumental in enabling me to overcome the challenges that I encountered. Despite his busy schedule, he made himself available to provide critical feedback and advice, even granting me the opportunity to serve as his teaching assistant.

I am equally grateful to my co-director, Foutse Khomh, whose diverse perspective on the research problem was extremely valuable in expanding my thinking beyond the research questions.

I would also like to extend my heartfelt gratitude to all the members of the MOOSE and SWAT research teams for their insightful contributions during our meetings.

Finally, I wish to express my appreciation to my esteemed committee members, Dr. Desmarais Michel and Dr. Hamdaqa Mohammad, for their willingness to evaluate my master's thesis.

RÉSUMÉ

L'importance de la perception de la qualité des applications par les utilisateurs est multiple et englobe plusieurs aspects critiques. Tout d'abord, il s'agit d'un retour d'information direct de la part des utilisateurs, qui reflète les performances de l'application par rapport à la description et aux promesses faites sur le magasin d'applications. Ce retour d'information est essentiel pour les développeurs, car il les aide à améliorer l'expérience de l'utilisateur. En outre, la perception de l'utilisateur joue un rôle essentiel dans la fidélisation des utilisateurs. Les clients satisfaits font preuve de loyauté à l'égard de l'application et sont enclins à la recommander à d'autres, ce qui permet d'élargir la base d'utilisateurs. En outre, une perception positive de l'utilisateur est en corrélation avec un meilleur classement dans les magasins d'applications, ce qui renforce la visibilité et la portée de l'application sur le marché numérique concurrentiel. Les recherches antérieures sur l'évaluation des applications mobiles à partir des critiques et des évaluations des applications ont généralement porté sur des applications de différentes catégories ou se sont concentrées sur un petit sous-ensemble d'applications d'une catégorie spécifique. Ces études proposent des conclusions générales pour l'ensemble d'un magasin d'applications comprenant des millions d'applications ou présentent des informations particulières pour un groupe limité d'applications. Néanmoins, une seule catégorie d'applications peut englober des dizaines, voire des centaines de milliers d'applications. Par exemple, AppBrain fait état de 46 625 applications dans la catégorie "Sports" des applications Google Play. D'autre part, une catégorie d'applications peut avoir des fonctionnalités très différentes (par exemple, les applications de paris sur le football et les applications d'entraînement au golf dans la catégorie "sports"). Il n'est pas clair comment les études de ces différents niveaux (c'est-à-dire le niveau du marché, le niveau catégoriel et le niveau fonctionnel) influencent les résultats de l'étude de la perception de la qualité des applications mobiles par les utilisateurs. C'est pourquoi cette thèse réalise une étude multi-niveaux à grande échelle des critiques et des évaluations des applications mobiles, en se concentrant sur les applications sportives Android. Cette thèse vise à comprendre comment les différents niveaux d'analyse des critiques et des évaluations des utilisateurs influencent les résultats de l'analyse. Pour atteindre cet objectif, la thèse est organisée autour de deux sous-objectifs.

Dans notre premier objectif, nous examinons des catégories d'applications ciblées, telles que les sports, qui peuvent fournir des informations plus précises que l'évaluation des applications dans toutes les catégories, tout en bénéficiant à de nombreux développeurs d'applications intéressés par cette catégorie. Cette étude vise à examiner de nombreuses applications au sein d'une même catégorie (les sports). Notre travail apporte une double contribution : 1) la

découverte d'informations spécifiques à des dizaines de milliers d'applications sportives et 2) l'apport de preuves empiriques sur les avantages de l'examen d'applications dans une catégorie particulière. Nous étudions empiriquement plus de deux mille applications sportives dans le Google Play Store, en examinant leurs caractéristiques (par exemple, les types de sports ciblés et les fonctionnalités principales) par le biais d'une analyse manuelle, les sujets des évaluations des utilisateurs par le biais de la modélisation des sujets, et les facteurs contribuant aux opinions négatives des utilisateurs par le biais de l'évaluation des utilisateurs et de l'analyse des sentiments. Nous avons identifié des applications sportives couvrant 16 types de sports (football, cricket, baseball, etc.) et 15 fonctionnalités principales (paris, conseils sur les paris, entraînement, suivi, etc.) En outre, nous avons extrait 14 thèmes des commentaires des utilisateurs, dont trois sont propres aux applications sportives (précision des prédictions, actualité et précision des outils). Nous avons conclu que les utilisateurs se plaignent principalement des publicités et de la qualité des applications sportives (par exemple, les bugs, la qualité du contenu, la qualité du streaming). L'analyse d'une catégorie spécifique d'applications (par exemple, les applications sportives) peut fournir des informations plus précises que l'examen d'applications de différentes catégories, tout en restant pertinente pour de nombreuses applications (par exemple, des dizaines de milliers). En outre, le marché des applications sportives, qui connaît une croissance rapide et une forte concurrence, offre de nombreuses possibilités de recherches futures, telles que l'étude de l'intégration de la science des données ou des techniques d'apprentissage automatique dans les applications logicielles ou l'examen des facteurs influençant la compétitivité des applications.

Dans notre deuxième objectif, nous analysons l'environnement extrêmement concurrentiel du magasin d'applications mobiles, qui compte un nombre impressionnant de 2,67 millions d'applications accessibles aux utilisateurs. Pour améliorer leur classement et s'assurer un avantage concurrentiel, les développeurs doivent comprendre les règles du jeu entre leurs rivaux. Nous proposons un système à trois niveaux (marché, catégorie et fonction) plutôt qu'une analyse globale pour aider les développeurs à comprendre les facteurs qui influencent les évaluations. Le marché correspond aux applications globales et générales du Google Play Store, la catégorie fait référence à un type spécifique d'application disponible dans le magasin d'applications (par exemple, les sports) et le niveau de fonction concerne les différentes fonctionnalités de la catégorie d'application spécifiée (par exemple, les paris sur le football pour la catégorie des sports). Dans cette étude, nous évaluons plus de 12 000 applications en utilisant une stratification des données à trois niveaux et nous examinons comment les caractéristiques de Google Play peuvent expliquer les évaluations des applications à différents niveaux. Notre analyse révèle que le niveau fonctionnel explique mieux les modèles que les niveaux supérieurs. En outre, notre approche multiniveaux nous permet d'identifier les car-

actéristiques dominantes propres à la catégorie. Pour l'analyse du niveau fonctionnel afin de comprendre les scores d'importance des caractéristiques, nous incorporons une étude basée sur le classement relatif, aidant les développeurs à identifier les caractéristiques cruciales pour surpasser leurs concurrents dans un paysage concurrentiel féroce. Nous concluons que la recherche sur l'approche multiniveaux peut aider les développeurs à mieux comprendre les facteurs uniques ayant un impact sur les évaluations des applications, ce qui leur permet d'identifier des paramètres spécifiques pour améliorer les évaluations par rapport à leurs concurrents.

ABSTRACT

The significance of user perception regarding app quality is multifaceted and encompasses several critical aspects. Foremost, it serves as direct feedback from users, reflecting the app’s performance in relation to the description and promises made on the app store. This feedback is instrumental for developers, as it guides them in enhancing the user experience. Additionally, user perception plays a pivotal role in user retention. Satisfied customers exhibit loyalty to the application and are inclined to recommend it to others, thereby expanding the user base. Moreover, positive user perception positively correlates with improved app store rankings, further improving the application’s visibility and reach within the competitive digital marketplace. Previous research on mobile app evaluations on app reviews and ratings typically examined apps across various categories or concentrated on a small subset of apps within a specific category. These studies either offer general findings for an entire app store comprising millions of apps or present particular insights for a limited group of apps. Nonetheless, a single app category can encompass tens or even hundreds of thousands of apps. For instance, AppBrain reports 46,625 apps in the “Sports” category of Google Play apps. On the other hand, a category of apps may have very different functionalities (e.g., football betting apps and golf training apps in the “sports” category). It is unclear how studies from these different dimensions (i.e., market dimension, categorical dimension, and functional dimension) impact the results of studying the user perception of the quality of mobile apps. Therefore, this thesis performs a large-scale multi-dimension study of the reviews and ratings of mobile apps, focusing on Android Sports apps. This thesis aims to understand how the different dimensions of analyses of user reviews and ratings impact the analysis results. To achieve this goal, the thesis is organized through two sub-objectives.

In our first objective, we examine targeted app categories, such as sports, which can yield more precise insights than evaluating apps across categories while still benefiting numerous app developers interested in that category. This objective aims to investigate a large number of apps within a single category (i.e., sports). Our work contributes in two ways: 1) uncovering insights specific to tens of thousands of sports apps and 2) providing empirical evidence on the advantages of examining apps in a particular category. We empirically investigate over two thousand sports apps in the Google Play Store, examining their characteristics (e.g., target sports types and primary functionalities) through manual analysis, user review topics via topic modeling, and factors contributing to negative user opinions through user rating and sentiment analysis. We identified sports apps covering 16 sports types (e.g., Football, Cricket, Baseball) and 15 primary functionalities (e.g., Betting, Betting Tips, Training,

Tracking). Furthermore, we extracted 14 topics from user reviews, three unique to sports apps (prediction accuracy, up-to-dateness, and tool precision). We concluded that users mainly complain about sports apps' advertisements and quality (e.g., bugs, content quality, streaming quality). Analyzing a specific app category (e.g., sports apps) can provide more precise insights than examining apps across various categories while remaining relevant for many apps (e.g., tens of thousands). Moreover, the rapidly-growing and competitive sports app market offers ample opportunities for future research, such as investigating the incorporation of data science or machine learning techniques in software applications or examining factors influencing app competitiveness.

In our second objective, we analyze the intensely competitive environment in the mobile app store, with an astounding 2.67 million apps accessible to users. To enhance their ratings and secure a competitive advantage, developers must comprehend the playing field among their rivals. We suggest a three-tier system (market, category, and function) rather than a global-dimension analysis to assist developers in understanding the factors affecting ratings. The market is the global and general apps in the Google Play Store, the category refers to a specific type of app available in the app store (e.g., sports), and the function dimension concerns the different functionalities of the specified app category (e.g., football betting in the sports category). In this study, we assess over 12,000 apps using the three-tier data stratification and explore how features from Google Play can explain app ratings at different dimensions. Our analysis reveals that the functional dimension modeling can better explain the app ratings than the category dimension. Similarly, category dimension explains better than market dimension. Additionally, our multidimension approach enables us to identify dominant features specific to the group of apps. In addition, for functional dimension analysis to identify feature importance scores among very closely rated applications, we incorporate relative ranking-based analysis, helping developers to identify crucial features to outperform competitors in a fiercely competitive landscape. We conclude that research on the multidimension approach can help developers better understand the unique factors impacting app ratings, allowing them to pinpoint specific metrics to improve ratings over their competitors.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
RÉSUMÉ	v
ABSTRACT	viii
TABLE OF CONTENTS	x
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
LIST OF SYMBOLS AND ACRONYMS	xv
CHAPTER 1 INTRODUCTION	1
1.1 Context and Motivation	1
1.2 Thesis Statement	2
1.3 Thesis Overview	3
1.4 Thesis Contributions	5
1.5 Thesis plan	6
CHAPTER 2 BACKGROUND	8
2.1 Mobile app analysis	8
2.1.1 Google Play Store	8
2.2 Mobile app review analysis	9
2.3 Mobile app rating analysis	9
2.4 Sports apps	10
CHAPTER 3 LITERATURE REVIEW	11
3.1 Prior works on app review analysis	11
3.2 Prior works on app rating rating analysis	12
3.3 Prior works on multidimension app data analysis	13
3.3.1 Prior works on sports applications	14

CHAPTER 4	A LARGE-SCALE EMPIRICAL STUDY OF ANDROID SPORTS APPS IN THE GOOGLE PLAY STORE	15
4.1	Introduction	15
4.2	Experiment Setup	18
4.2.1	Identifying the Sports Apps	18
4.2.2	Manual analysis	20
4.2.3	Collecting reviews	20
4.2.4	Topic Modeling	21
4.2.5	Sentiment analysis of user reviews	21
4.3	Experiment Results	21
4.3.1	RQ1: What are the characteristics of sports apps?	21
4.3.2	RQ2: What are the topics raised by sports apps users in their reviews?	29
4.3.3	RQ3: What do users complain about sports apps?	36
4.4	Discussions	42
4.5	Threats to Validity	44
4.5.1	Internal Validity	45
4.5.2	External Validity	45
4.5.3	Construct Validity	45
4.6	Chapter Summary	46
CHAPTER 5	HOW CAN WE EXPLAIN USER RATINGS OF MOBILE APPS? A MULTI-DIMENSION STUDY OF ANDROID SPORTS APPS.	47
5.1	Introduction	47
5.2	Experiment Setup	51
5.2.1	A: Data Gathering	51
5.2.2	B: Data Cleaning	57
5.2.3	C: Data Stratification	58
5.2.4	E: Rating prediction models	60
5.2.5	D: Important feature identification	62
5.2.6	F: Relative ranking prediction	62
5.3	Experiment Results	63
5.3.1	RQ1:How do regression models perform in predicting rating across three dimensions?	63
5.3.2	RQ2: What are the most influential features for predicting ratings for each dimension?	69
5.3.3	RQ3: What factor influence the ranking of apps among competitors?	73

5.4	Discussions	76
5.5	Threats to validity	78
5.5.1	Internal Validity	78
5.5.2	External Validity	78
5.5.3	Construct Validity	79
5.6	Chapter summary	79
CHAPTER 6 CONCLUSION		81
6.1	Summary	81
6.2	Limitations of the proposed approaches	83
6.3	Future research	84
REFERENCES		85

LIST OF TABLES

Table 4.1	Cumulative number of reviews	21
Table 4.2	The distribution of the number of downloads	25
Table 4.3	Characteristics of the manual analyzed sports apps	26
Table 4.4	Topics extracted from the reviews of sports apps.	32
Table 4.5	Average sentiment and average rating score of reviews	42
Table 4.6	Distribution of average review sentiment over different topics	43
Table 5.1	The extracted app features from Google Play Store	53
Table 5.3	Rating prediction model performance	67
Table 5.4	Model performance across different dimensions: Common testset	68
Table 5.5	Feature ranking using Scott Knott	71
Table 5.6	Model performance on relative ranking in functional dimension	74
Table 5.7	Feature importance among competitors	75

LIST OF FIGURES

Figure 4.1	Overview of our study	19
Figure 4.2	The number of sports apps launched per year in Appbrain's top-500 list.	24
Figure 4.3	The distribution of rating scores of the studied sports apps.	24
Figure 4.4	Distribution of reviews over different functionalities and topics.	36
Figure 4.5	Pairwise Spearman correlation between Average sentiment, Percentage of negative reviews, Average rating score and Percentage of negative rating score of the reviews of each apps(after removing outliers). The * symbols along side the correlation values indicate the statistical significance of the respective correlation: No symbol $p \geq 0.05$, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$	40
Figure 4.6	Distribution of review sentiment over user rating associated with reviews (before reviewing the outliers).	41
Figure 5.1	Overview of our study	52
Figure 5.2	Distribution of different functionalities in sports category	59
Figure 5.3	Distribution of average app rating for each data dimension. Each box-plot shows the distribution of the average rating of the all apps in the given dimension	64

LIST OF SYMBOLS AND ACRONYMS

LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
RFG	Random Forest Regressor
BR	Bagging Regressor
DR	Dummy Regressor
XGB	XG Boost Regressor
CBR	Case-Based Reasoning
MSE	Mean Squared Error
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
NRMSE	Normalised Root Mean Squared Error
SD	Standard Deviation
IQR	InterQuartile Range
ESD	Effective Size Difference

CHAPTER 1 INTRODUCTION

1.1 Context and Motivation

The mobile app industry has experienced significant growth in recent years, driven by increased smartphone adoption, advancements in 5G technology, and artificial intelligence [1]. Mobile apps span a range of genres, including Games, Education, Business, Personalization, and Sports. Data analysis is crucial to any mobile app development strategy. It provides insights that help developers understand user behavior, identify trends, and make informed decisions contribute to their mobile app's success.

Popular app markets include the Google Play Store for Android devices, the Apple App Store for iOS devices, and the Samsung Galaxy Store for Samsung devices. Each market has unique features and audiences, making it important for developers to consider their target users and platform-specific guidelines when developing and launching their mobile apps.

The direct feedback users provide through app markets allows developers to gain insight into user perception of app quality, including reviews and ratings. The importance of app reviews and ratings is evident across different facets. Developers can utilize this feedback as a primary source for product quality analysis and feature improvement [2]. Moreover, app ratings and reviews create a sense of trust among users and can influence their decision to download an app, ultimately increasing user retention for specific applications. Understanding app ratings and reviews trends is crucial in maintaining user trust and satisfaction [3, 4]. The higher the user retention and app quality, the better the application's ranking in the app stores, ultimately leading to greater app success [5]. Thus, user perception of app quality is vital for the success of any mobile app.

Previous research has attempted to understand user perception of app quality through app reviews [6, 7, 8] and app rating [9, 10, 11]. However, much of the research focuses on either a global-dimension evaluation, which can be too generic, or on a small set of specific groups, which may not capture the full picture. It is crucial to analyze the user-perceived qualities of applications at both the global and competitor-specific dimensions to understand the granularity and identify unique important factors. To overcome these limitations, it is crucial to analyze the user-perceived qualities of applications at both the global and competitor-specific dimensions. This approach can provide a more granular understanding of user perception of app quality and help identify the unique important factors that contribute to user satisfaction. By understanding these key factors, developers can make informed decisions about

feature development and user retention, ultimately contributing to the success of their mobile applications.

In our thesis, we propose a three-dimension approach to analyzing mobile apps. This approach includes the Market dimension, the Category dimension, and the Functional dimension. At the Market dimension, we analyze global apps from the app store, which provides a broad overview of the app market. At the Category dimension, we analyze genres provided by the app store or apps under specific keywords spanned across multiple genres, which provides a more specific view of the market(e.g., sports apps, personalization apps). Finally, at the Functional dimension, we analyze applications that propose similar functionality to the user or competitor apps, which provides the most granular view of the market(e.g., Football Betting apps, Fitness Training apps). By analyzing data at three different dimensions, our approach allows for a more nuanced understanding of the unique characteristics of mobile apps. This enables us to identify important factors contributing to user satisfaction and develop effective strategies for improving app quality.

For our thesis, we are considering Sports apps as our case study subject. Sports apps are among the top categories of mobile apps with high user engagement and fast-paced growth, with consistent and niche fan bases [12]. These apps contain a variety of functionalities, including social media, personalization, live streaming, and betting. To ensure the success of a sports app, it is essential to consider user-perceived information to improve its performance. User reviews and ratings are direct feedback from users to developers, providing insight into the app’s quality and usability.

1.2 Thesis Statement

In this research, we analyze the significance of user-perceived quality of data in supporting mobile application developers and facilitating their success. We highlight that studying the user perception of mobile app quality can be impacted by the dimension of studies (market dimension, categorical dimension, functional dimension). With the exponential growth of mobile app development, utilizing a multidimension analysis approach has become paramount to studying the user-perceived quality of mobile apps. Mobile app genres typically comprise over ten thousand apps, each comprising diverse functionalities. An overarching analysis may not provide granular insights into the perceived quality of apps compared to their competitors or apps with similar functionality. Hence, sub-dimension analyses such as categorical and functional analyses are necessary to understand the users’ perceptions of app quality. We conduct two empirical studies to address this need, leveraging large-scale app data at multiple dimensions. The first study will focus on the reviews (qualitative information on users’

perceived quality of apps) to understand what are the topics being discussed and how users' feel about them. For the second study, we will focus on rating scores (quantitative information on users' perceived quality of apps) to understand the features affecting the rating scores. Our case study focuses on sports applications, examining over 10,000 apps at the market dimension, more than 1,000 at the categorical dimension (specifically, sports apps), and over 100 apps within each functional dimension of the sports app category. For both of our studies, we compare and conclude how each dimension complements the other in terms of information that we can conclude for the app quality.

Comprehending the different degrees of users' perception of app quality can aid in finding unique trends and patterns that can be utilized to improve the app quality. This research thesis constitutes an initial attempt at a multidimension large-scale examination of the user-perceived quality of app data in the form of reviews and ratings to point out differences in the information that can be concluded at the market, categorical and functional dimensions using sports apps as a case study. Its purpose is to contribute to the realm of mobile app analysis by concentrating specifically on the extracted and curated user-perceived quality of data (reviews and ratings) in the domain of sports apps. Through the utilization of statistical analysis (trend analysis, correlation analysis and redundancy analysis) and machine learning approaches (topic modelling, sentimental analysis and regression models) inspired by previous studies, we aim to discern crucial insights and trends that may be employed to elevate the design, functionality, and overall user experience of sports applications among competitors (functional dimension) within the sports app genre (categorical dimension), and in the Google Play Store (market dimension).

1.3 Thesis Overview

This thesis presents two multidimension empirical studies focused on analyzing the user-perceived quality of app data where sports apps as a case study. The first study is a large-scale empirical investigation of the reviews of Android sports apps in the Google Play Store. We collect data directly from the Play Store and analyze a large number of sports apps to identify unique insights specific to the sports app category. We extract and manually classify the data and then perform a comprehensive analysis of user reviews using topic modeling to identify common themes and issues users raise. We also apply sentiment analysis to understand the factors that users complain about, providing valuable insights into user perception of sports apps' quality and user experience. We analyze the differences in the complaining factors at functional, meaning a specific functionality of apps(i.e., competitors) and categorical dimensions (i.e., genre sports) and compare the results to existing works at

market dimensions.

The second study analyzes the rating scores of Android apps in the Google Play store, where we selected the scope of our study in a categorical dimension of sports apps and analyzed them at three dimensions: market, category, and function. We leverage this multidimension analysis to explore differences in rating score patterns and identify important features across the different dimensions of analysis. By comprehensively analyzing large-scale data, we aim to provide a more detailed and nuanced understanding of sports app user behaviour and perception. Our research contributes to the growing body of literature on mobile app analysis and provides practical insights for app developers looking to improve the quality of their sports apps.

In the following, we elaborate on each of the studies:

1. *An empirical study of android sports apps in the Google Play Store:* This research study aims to investigate the perceived quality of mobile applications by examining user app reviews in the Google Play Store, utilizing a multidimension analysis approach. Specifically, the study focuses on sports apps as its chosen case study. Through empirical research, this study seeks to gain a deeper understanding of how users perceive the quality of mobile applications. In the study, we identify the functionalities, types, and algorithms we collected in sports apps. To achieve this, we developed a unique keyword-based search algorithm to collect and curate many sports apps, which we manually labelled to gain a comprehensive understanding of their characteristics. Our analysis of the distribution of sports apps led us to identify the most prevalent types and functionalities in sports software development and the extent to which predictive algorithms are used in this field. In a subsequent investigation, we proposed applying a topic modelling algorithm and examining user-perceived reviews to gain insights into users' feedback. Our findings revealed 14 distinct topics, three unique topics to sports apps. Furthermore, we conducted a sentiment analysis to identify user concerns expressed in the reviews, with a view to improving user experience and quality. Finally, we compared our results with those of general app analysis(Market dimension study), specific genre(e.g., sports that is categorical dimension) and a specific functionality we identified(e.g., football betting apps that is functional dimension). We discovered that a category-specific study enabled us to uncover four new topics that had not been previously discussed, thus contributing unique insights to the field. Similarly, a functional-dimension study identifies unique user sentiment associated with the topic for the given functionality.

2. *How can we explain user rating of mobile apps:* This research paper focuses on the critical role of user ratings as a key source of user-perceived information. Specifically, we aim to analyze the significance of employing a multidimension analysis to uncover the

relationship between user rating patterns and their underlying factors. Our three-dimension study (market, category, and functional) demonstrated that each dimension should be examined separately to yield unique insights. We also found that converting ratings to relative rankings is necessary when the ratings are exceptionally high, enabling a more detailed analysis of the factors affecting user ratings. Our study further showed that rating predictions are more effective at lower dimensions and are useful for identifying functional-specific, category-specific, and general features. The functional-specific and category-specific features are not identified at the market dimensions analysis. In multidimension analysis, the lower dimension has more intricate details about the specific dimension, which helps to locate the specific information that is affecting the rating score in the given dimension. Lastly, the relative ranking mechanism can be used to identify critical features when competition is high, where absolute ratings are very close and will not be able to locate the important features.

1.4 Thesis Contributions

This thesis empirically investigates large-scale sports application data acquired from the Google Play Store. The primary aim of this research is to scrutinize multidimension user-perceived information, encompassing user reviews and ratings, to discern valuable insights that can aid developers in augmenting the user experience.

- We have introduced an iterative keyword-based approach for extracting apps from the Google Play Store that belong to a specific category. This method ensures greater inclusivity of apps that are spread across different genres but are still relevant to the category in question.
- Our research revealed the identification of 14 distinct types of sports covered by the apps, 13 diverse functionalities of sports covered by the apps, and the use of four distinct analytical/statistical algorithms being used in sports apps. (e.g., *Football-Betting Tips-Predictive algorithm* here Football is the sports type, Betting Tips is the functionality among the sports type, and Predictive Algorithm is the algorithm being used in the given apps).
- Through the utilization of topic modelling techniques, we have identified 14 topics that are the subject of discussion within the reviews of sports applications. The subsequent manual analysis of these topics enabled a review and analysis of their names. Notably, three of these topics are exclusive to sports apps, namely, the *accuracy of prediction*, *up-to-dateness*, and *precision of tools*. This outcome provides a compelling illustration of how an analytical approach that is focused on a specific category can furnish a unique

and relevant insight that is both meaningful and valuable for that particular category of application.

- Through sentiment analysis of user reviews, we identified topics that users express dissatisfaction with. On a functional dimension analysis, For *News apps*, *Streaming quality* and *Advertisement* are the most concerning topics, which is different from *League management apps* where *Quality of content* and *Version updates* are the concerned topic.
- We found that *Streaming*, *Betting*, *Team management*, and *League management* are the functions users most frequently complain about.
- We identified that *categorical dimension* analysis of app ratings performs better than a *market-dimension analysis*, while a *functional dimension* analysis further improves the performance. And the model that performed best is consistent over all the dimensions(i.e., Random Forest regressor).
- We observed that multidimension analysis helps to identify category-specific important features that are related to the ratings, e.g., our analysis has revealed that the *precise location (GPS and network-based)* , *run at startup*, *videoImage*, *pair with Bluetooth device* is an important factor for the *Sports* category at the Category dimension. However, this factor was not identified in the Market dimension analysis.
- We observed that in functional dimension analysis using relative ranking could be particularly helpful when the competition landscape is limited, as it can help identify unique metrics affecting the rating. In cases where the number of competing apps is small, the global analysis rating analysis may not provide enough granularity to identify key factors influencing user satisfaction. e.g., *permission to draw over other apps* for *Football- Streaming apps* identified in the relative ranking-based analysis on the functional dimension, which was not identified at a global dimension or absolute rating-based analysis and very unique to streaming applications.

1.5 Thesis plan

The remainder of the thesis is organized as follows. Chapter 2 introduces the fundamental concepts related to mobile and sports app development challenges. Chapter 3 discusses the related works in the domain of mobile app data analysis and sports app data analysis. In Chapter 4, presents our first empirical study, in which we examine sports apps and their reviews. Chapter 5 Presents our second empirical study where we conduct a multidimension

sports analysis on their app rating. Chapter 6 concludes the thesis and discusses some avenues for future works.

CHAPTER 2 BACKGROUND

This thesis investigates app quality by analyzing user-perceived information on a large scale, such as sports app reviews and ratings. This chapter introduces the background knowledge of mobile app analysis, the Google Play Store, mobile app review analysis, mobile app rating analysis, and sports apps. This information will be useful in comprehending the subsequent chapters of this thesis.

2.1 Mobile app analysis

The mobile app industry has experienced significant growth in recent years, with a projected worth of approximately 1 billion USD. In 2008, Apple’s introduction of the App Store provided developers with a platform to distribute and sell their apps. Following the App Store’s success, similar app marketplaces emerged, such as the Google Play Store and Samsung store [13]. Studies have shown that the concept of crowd-sourced opinion is relatively new, offering users immediate access to information that can influence their purchasing decisions [14]. Consequently, developers have shifted their focus toward enhancing user experience in addition to improving technical aspects, leading to the development of app store optimization (ASO) and the increasing use of user reviews and ratings as a source of feedback and evaluation, which have both contributed to the evolution of mobile app analysis. Today, mobile app analysis involves a broad range of techniques and tools, including app store data analysis (Data.ai [15], Sensor Tower [16], Mobile Action [17]), user review analysis (App bot [18], crash and error analysis (Firebase crashlytics [19]), user behaviour analysis (Mixpanel [20]), and more. Mobile app analysis is essential in helping developers and marketers understand their app’s performance and user perception, allowing them to make data-driven decisions to improve their app’s performance and user experience [21].

2.1.1 Google Play Store

The Android user base has experienced exponential growth since 2009, solidifying its status as one of the largest mobile app software platforms available. The primary distribution channel for Android apps is the Google Play Store, which currently accounts for about 70% of the global smartphone market share [22]. The Google Play Store has a vast selection of apps that has grown considerably over the years, with users able to access apps for various purposes such as communication, productivity, entertainment, and education [23]. Currently, there are

over 2.67 million applications available for download, with accompanying metadata such as user reviews and ratings. Additionally, the Android app development community has grown considerably, with active discussions and support available both from developers and Google itself. Google Play Developer Console offers numerous tools for app developers to track and analyze their app’s performance, providing valuable information on user acquisition, retention, engagement, and revenue generation [24]. However, given Google’s unique ranking and maintenance practices with protective measures in place [25], it is crucial to analyze the available data to aid developers in making their application visible and successful in the competitive marketplace.

2.2 Mobile app review analysis

During the early stages of app market growth, such as in 2009 when Google Play store was launched, developers started to receive direct user reviews and began analyzing them. However, at that time, the tools and approaches available were limited, and the studies focused mainly on keyword-based analysis. With the advancement of software engineering practices over the years, mobile app review analysis has also progressed. As the app market matured, businesses and developers recognized the importance of user feedback in improving app quality, driving user engagement, and increasing revenue. This realization led to the emergence of app analytics tools such as Data.ai (formerly known as App Annie) [15] and Sensor Tower [16]. These tools provided basic analysis of app store data, including user reviews and ratings. With the advent of machine learning technologies, mobile app review analysis has become even more sophisticated. Techniques such as Natural Language Processing (NLP), Sentiment Analysis, and Topic Modelling have been integrated into app analytics platforms, enabling developers and businesses to understand user sentiment and identify areas for improvement quickly [3, 6, 26]. These techniques have proven to be highly effective in retrieving more information about the reviews. As the benefits of mobile app review analysis became apparent, businesses and developers started incorporating review analysis insights into their app development and business processes. This includes using feedback to prioritize feature development, address user concerns, and inform marketing strategies [27].

2.3 Mobile app rating analysis

The evolution of mobile app rating analysis and reviews has exhibited a significant growth pattern in recent years. Initially, app ratings were merely aggregated, and industry analytical tools for review analysis were also utilized for rating analysis. However, with the increasing

competition in the mobile app market, the demand for sophisticated mobile app rating analysis tools has increased. The available tools range from simple rating aggregation to advanced sentiment analysis and machine learning algorithms. Popular examples of such tools include Data.ai (formerly known as App Annie) [15], which offers app rating and review analysis and app store optimization tools, and Sensor Tower [16], which provides app rating and review analysis, market intelligence, and advertising analytics. The significance of mobile app rating analysis for developers and businesses striving to remain competitive in the expanding mobile app market is only expected to grow.

2.4 Sports apps

The need for sports organizations, leagues, and fans to better understand athlete and team performance, as well as fan preferences and behaviors, has led to the development of sports apps. Sports app analysis facilitates the optimization of operations for clubs and leagues, supports data-driven decision-making, and allows for more personalized fan experiences [28]. For example, the digitalization of sports betting has enabled increased monetization and revenue through sports apps. Popular examples of sports betting apps include FanDuel, which offers daily fantasy sports and traditional sports betting options, and Bet365, which is a widely recognized sports betting app that provides an extensive range of sports betting options, including live in-game betting. Additionally, sports apps include a wide variety of options, including social media and streaming. Seasonal trends play a significant role in sports app usage and engagement, contributing to the consistent growth of the industry [29]

CHAPTER 3 LITERATURE REVIEW

Investigating user perception of mobile app quality in a large-scale multidimension setting is the goal of my thesis. Specifically, we'll examine ratings and reviews as our primary source for understanding data from users' perspectives. Thus, this section first discusses existing literature on app rating and review analyses. In addition, we study prior work on multidimension app analysis (relevant to our methodology) and that on sports app analysis (relevant to our studied subject apps).

3.1 Prior works on app review analysis

In this section, our attention is directed toward analyzing mobile app reviews. This category is divided into two parts: firstly, the reviews are characterized by examining the various motivations behind conducting app review analyses, and secondly, the methodologies utilized in analyzing the reviews are reviewed to gain a deeper understanding of the practices involved.

Characterizing review analysis: Many of these studies were motivated by user opinions or complaints [6, 30, 31, 32, 33, 34]. For instance, Fu et al. (2013) [6] present a study with sentiment analysis and topic modelling that sheds light on the reasons why users give negative feedback in mobile app stores. The authors propose a framework for analyzing negative feedback and identifying common themes, which can help app developers understand the reasons behind negative feedback and prioritize improvements. Chen et al. (2014) [35] aim to help mobile app developers identify informative reviews from the large volume of user feedback in mobile app marketplaces. The authors argue that informative reviews contain valuable feedback that can be used to improve apps, but it can be difficult for developers to identify them. Later the same year, Wano and Iio (2014) [36] presented a manual analysis of the relationship between the app genre and the reviews they receive in mobile app stores. The results demonstrate that certain genres tend to receive more reviews and that the sentiment of these reviews varies across categories. Researchers have conducted several studies in the past few years to comprehend the user-perceived quality of apps in reviews with specific functionalities, such as E-scooter, Drones, and Covid-19 apps [3, 37, 38]. These studies have indicated that analyzing user reviews can give developers and researchers valuable insights, as app reviews often contain more detailed and specific information than generic analyses. Such insights can assist developers in improving the quality of their apps and enable researchers to understand users' needs and preferences better.

These studies demonstrate that reviews can provide valuable insights into user concerns and improve app quality. However, previous research has primarily analyzed app reviews from a general perspective across various categories or focused on a small number of apps. This thesis highlights the potential benefits of analyzing reviews for a specific app category, such as sports, which often includes many apps. It compares the results across the market dimension, specific genre and specific functionality of sports apps in the Google play store.

Practices in review analysis: In the early stages of mobile app review analysis, developers recognized the importance of incorporating user feedback into the app development life cycle. Various methods were employed to classify and understand user reviews, including Manual analysis [39, 40], keyword-based analysis and rule-based approaches [31, 41]. However, the emergence of machine learning-based techniques marked a significant shift in review analysis [6, 7, 26, 30, 32, 33, 42]. While manual analysis of mobile app reviews is generally considered more accurate, it can be difficult to scale for large datasets. In such cases, machine learning approaches can offer a more practical alternative with relatively high accuracy [43]. Topic modelling has been widely adopted for understanding similar topics that is being discussed in reviews [3, 6, 42, 44, 45, 46, 47]. Considering the scholarly approaches using textual and non-textual approaches or supervised machine learning approaches, topic modelling is the most common practice followed on the large-scale unsupervised grouping of reviews [43, 48]. There has been a sentiment analysis technique to support and analyze the topics identified [3, 6, 49]

Previous studies indicate that topic modelling techniques can be helpful for analyzing large-scale unsupervised data to understand the user perception of the quality of apps, with the support of additional machine learning or statistical analysis techniques to support the results.

3.2 Prior works on app rating rating analysis

In this section, our focus is on analyzing mobile apps ratings. We have divided this category into two sections: firstly, we characterize the app ratings by examining the various motivations behind rating an app, and secondly, we review the different methodologies used in analyzing the ratings to gain a deeper understanding of the practices involved.

Characterizing rating analysis: Other studies have attempted to identify the statistical relationship between app ratings and a minimal number of features [50, 51]. However, all rating prediction methods are constrained by the data available in the specific app store being analyzed and the features under consideration in the research. Studies on app store optimiza-

tion factors have underscored the significance of metadata, such as app name, description, and visual elements, which are available in the app store [5]. Furthermore, some studies have investigated the association between app ranking and ratings, assuming that ratings are a crucial factor that affects app ranking [5]. A limited number of studies have compared the relationship between ratings and reviews [6, 52]. A market-dimension analysis on influencing features for rating predictions has been done [10, 11]. However, identifying areas for improvement for app developers to improve app ratings within the context of the genre (category dimension) and competition (functional dimension) remains an open research question.

Previous research has analyzed rating prediction models or rating analysis for mobile apps, either at a market dimension or limited to one specific genre. Some studies have sought to identify feature importance for market-dimension prediction or prediction for individual applications. These studies have concluded that rating prediction models or analysing relationships with a rating and other features can provide valuable insights for app improvement.

Practices in rating analysis: The evolution of application rating prediction has advanced from simple correlation and statistical analysis [53, 54, 55] to more sophisticated machine learning algorithms [9, 10, 11, 56, 57]. For example, Sarro et al. [9] proposed a Case-Based Reasoning model based on the features described in the app description to predict app rating. In contrast, Gnotthivongsa and Huang [56] utilized Collective Matrix Factorization. Umer et al. [11] conduct a study on predicting numeric ratings for Google apps using text features and ensemble learning models for regression. In their study, Mahmood et al. [10] adopted a similar approach to regression learning models.

Prior research has suggested that regression models perform better for the given dataset among the benchmark models investigated (e.g., Random Guess [9]). However, as our study focuses on identifying the features that influence app ratings on a large-scale, multidimension dataset, we also prioritize the models' explainability. Therefore, we limit our selection of models to case-based reasoning and ensemble learning.

3.3 Prior works on multidimension app data analysis

Multidimension analysis is a powerful tool for gaining insights into complex phenomena. Although no prior studies have explored multidimension mobile app analysis for rating, research has demonstrated that comparing characteristics amongst the competitors (functional dimension) can provide more intuitive information. Google Play Store itself introduced a new feature called *custom peer group* where developers can compare their app's Android vitals and ratings data with a group of apps that they select [58]. This triggered the research spectrum

to analyze apps within a peer app analysis (functional dimension) [26, 59, 60]. Hassan et al. [26] did an analysis of peer apps that can offer a distinctive and more suitable perspective compared to a global analysis of apps. This is because a specific topic in app reviews that may be crucial for one peer group may not hold the same dimension of importance for other peer groups or when examined from a global perspective.

In the context of review analysis Fu et al. [6] analyzed user feedback in a mobile app store to understand why users give negative feedback to certain apps. They used a multidimension (app, user, and feedback dimensions) approach to analyze the data, which allowed them to identify patterns and trends across multiple dimensions of analysis. The authors found that using a multidimension approach enabled them to understand user feedback and its underlying factors better. Similarly, Noei et al. [42] investigated the challenge of analyzing a large number of user reviews in mobile app stores; they emphasize in their study that functional dimension analysis brings more context to comprehend the user reviews.

According to previous research, analyzing an app’s functionality in detail can reveal more insights than looking at it from a broader perspective. Therefore, to better understand how users perceive an app’s quality, it’s crucial to conduct a multidimension analysis that can uncover unique information that pertains specifically to that dimension.

3.3.1 Prior works on sports applications

Some prior works on mobile applications have focused their studies on sports apps. A study evaluates the quality of recreational sports apps [61] by evaluating the apps using an expert panel approach. Francois et al. [62] identified the low quality of free coaching apps by considering 30 apps. The content of these apps was compared against the current guidelines and fitness principles established by the American College of Sports Medicine (ACSM). Pengcheng et al. [63] evaluated the need for Sports and Fitness Apps from the Perspective of “Healthy China”. The paper investigates the development status of sports and fitness apps and user satisfaction. There are further studies on the qualities of the apps. Francois et al. [64] evaluated the fitness apps’ quality with two independent raters. Another study on analyzing the potential of the fitness apps from Adria et al. [65] evaluates the validity and reliability of apps designed for CRF assessment. Despite being one of the most widely used categories of mobile apps [23], there is a dearth of research specifically focused on sports as a category of apps. Given the wide range of functionalities encompassed within this category, it presents an excellent opportunity for a comprehensive case study.

CHAPTER 4 A LARGE-SCALE EMPIRICAL STUDY OF ANDROID SPORTS APPS IN THE GOOGLE PLAY STORE

Prior studies on mobile app analysis often analyze apps across different categories or focus on a small set of apps within a category. These studies either provide general insights for an entire app store which consists of millions of apps, or provide specific insights for a small set of apps. However, a single app category can often contain tens of thousands to hundreds of thousands of apps. For example, according to AppBrain, there are 46,625 apps in the “Sports” category of Google Play apps. Analyzing such a targeted category of apps can provide more specific insights than analyzing apps across categories while still benefiting many app developers interested in the category. This work aims to study a large number of apps from a single category (i.e., the *sports* category). Our work can provide two folds contributions: 1) identifying insights that are specific to tens of thousands of sports apps, and 2) providing empirical evidence on the benefits of analyzing apps in a specific category.

We perform an empirical study on over two thousand sports apps in the Google Play Store. We study the characteristics of these apps (e.g., their targeted sports types and main functionalities) through manual analysis, the topics in the user review through topic modeling, as well as the aspects that contribute to the negative opinions of users through analysis of user ratings and sentiment. We identified sports apps that cover 16 sports types (e.g., Football, Cricket, Baseball) and 15 main functionalities (e.g., Betting, Betting Tips, Training, Tracking). We also extracted 14 topics from the user reviews, among which three are specific to sports apps (*accuracy of prediction*, *up-to-dateness*, and *precision of tools*). Finally, we observed that users are mainly complaining about the advertisements and quality (e.g., bugs, content quality, streaming quality) of sports apps. It is concluded that analyzing a targeted category of apps (e.g., sports apps) can provide more specific insights than analyzing apps across different categories while still being relevant for a large number (e.g., tens of thousands) of apps. Besides, as a rapid-growing and competitive market, sports apps provide rich opportunities for future research, for example, to study the integration of data science or machine learning techniques in software applications or to study the factors that influence the competitiveness of the apps.

4.1 Introduction

Mobile app stores (e.g., Google Play Store) provide a rich collection of apps for users to download and a wealth of information concerning users’ feedback on the apps (e.g., user

reviews). Prior work has performed extensive studies on mobile app information that is available in app stores [66]. These studies often analyze a large number of apps across different app categories (e.g., [67], [26], [30] or focus on a small set of particular apps (e.g., [30], [68], [69]). For example, Chen et al. [67] collected the reviews of 31,518 apps across all app categories of the Google Play Store (including the top 500 popular apps from each category), identified UI-related reviews, then manually analyzed a random sample of UI-related reviews to characterize UI-related issues. The findings of these types of studies are usually generalizable to different app categories. However, such findings may miss the particularity of a specific app category. On the other hand, studies that focus on a small set of apps may derive findings that are too specific for the studied apps and lack generalizability.

An app category in the app store usually contains thousands to hundreds of thousands of apps. For example, there are 46,625 apps in the *sports* category in the Google Play Store [1]. We assume that analyzing apps from such a targeted category can provide more *specific* insights (i.e., specific to the app category) for *a large number* of apps (e.g., tens of thousands of apps) in the app store. Thus, in this work, we aim to validate our assumption and perform a study of a large number of Android apps from a single category. Specifically, we focus on the apps in the *sports* category in the Google Play Store.

We focus our study on the *sports* app category for several reasons. First, it is a large and fast-growing app category. It is reported that the *sports* category includes over 46K apps [1]. In recent years, the *sports* category has been growing very fast, with 35% of the top 500 *sports* apps released in the last three years.

It is predicted that the sports technology industry is going to have an average annual growth of 16.8% and reach a market value of \$36.2 billion USD by 2028, and smartphone apps make up a key component in this industry [70]. Thus, we expect that *sports* apps will continue to grow rapidly for at least a few years to come. Second, although apps in the *sports* category share some similarities, they cover a variety of sports domains (e.g., soccer or basketball) and functionalities (e.g., training or news), and widely adopt emerging technologies such as data analytics, artificial intelligence, internet of things, and social media integration [70]. Such variety enables our study to identify common issues across the entire app category, the ones that are specific to a sub-category of apps.

In this work, we study a carefully curated set of 2,058 *sports* apps in the Google Play Store. As far as we know, no prior work has studied such a large number of mobile apps in a single app category. We exclude *sports game* apps as they have very different characteristics from our considered *sports* apps. Based on the top 500 free sports apps provided by AppBrain [71], we extracted keywords related to *sports* apps, then searched through the Google Play Store

and manually verified each of the resulting apps. Our work combines qualitative analysis and quantitative analysis to understand the characteristics of the *sports* apps, the topics in users’ reviews, and the factors that users complain about these apps. Our work highlights the benefits of analyzing apps in a specific category and provides insights for *sports* app developers to improve their apps or build new apps. Specifically, our work aims to answer the following three research questions (RQs).

- **RQ1: What are the characteristics of *sports* apps?** To understand the characteristics of sports apps, we manually study a statistically representative sample of sports apps. We observe that sports apps cover a wide spectrum of sports types (e.g., football) and functionalities (e.g., training). On the other hand, the sports app market is highly competitive, with many apps covering the same sports types and functionalities.
- **RQ2: What are the topics raised by *sports* app users in their reviews?** Understanding users’ perceptions of the apps communicated in their reviews can aid in validating the quality of the apps [72, 73]. In this research question, we apply topic modeling to extract topics from the user reviews of sports apps. We derived 14 topics from the users’ reviews of the 2,058 studied sports apps, including topics identified in previous review analysis studies, as well as three new topics that are specific for sports apps including *accuracy of prediction*, *up-to-dateness*, and *precision of tools*.
- **RQ3: What do users complain most about *sports* apps?** User ratings and reviews are indicators of how users feel about the apps. In this research question, we analyze the user ratings and the sentiment of the user reviews to understand the factors that users complain about sports apps. We observe that users are mainly complaining about the advertisements and quality (bugs, content quality, streaming quality) of the apps. Besides, *Streaming*, *Betting*, *Team management* and *League management* are the most complained about sports app functionalities.

Through our study of the 2,058 apps in the sports app category, we conclude that analyzing a targeted category of apps (e.g., sports apps) can provide more specific insights than analyzing apps across different categories while still being relevant for a large number (e.g., tens of thousands) of apps. For example, we identified more specific characteristics of the apps (e.g., sports types, functionalities) and new topics raised by users in their reviews that are specific to sports apps. In addition, we observe that sports apps are a rapid-growing and competitive market, which provides rich opportunities for future research, for example, to study the integration of data science or machine learning techniques in software applications, or to study the factors that influence the competitiveness of the apps in the competitive market.

To help future work replicate or extend our work, we share our replication package¹.

Chapter organization. The remainder of the chapter is organized as follows. Section 4.2 describes the experimental setup of our study. Section 4.3 describes our results from the experiments conducted for each research question. Section 4.4 discusses the implications of our findings. The threats to the validity of our findings are discussed in section 4.5. Finally, we conclude our chapter in section 4.6.

4.2 Experiment Setup

Figure 5.1 shows an overview of our study. This section explains our process of identifying, collecting, and preparing the data for answering our research questions. The detailed approaches used for answering each RQ are described in Section 4.3.

4.2.1 Identifying the Sports Apps

Google Play store provided category tags for each app. However, from our preliminary analysis, we understood that developers placing their apps in the play store do not pay much attention to the tags available. Thus the category tags available in the play store are not reliable [74]. Besides, the Google Play store only returns a limited number of apps for each search query (e.g., searching the keyword “sports” return fewer than 250 apps). To identify sports apps from all the other apps, we need to have a set of keywords that we can use as the search keywords. To address these challenges and identify a large number of *sports* apps, we started with an initial set of *sports* apps, based on which we identified a set of related keywords. Then, we searched these keywords in the Google Play store and finally manually verified each of the resulting apps. The approach we used to generate those keywords is to collate the description associated with the apps and thus look into the contextually important words to use as the search words for the next iteration.

Top-rated sports apps from AppBrain

AppBrain² is one of the leading companies that collect information regarding the Android app ecosystem. From AppBrain, we downloaded a dataset of top-rated 500 free apps [71]³ in the US region. From each of these apps, we extracted the following metadata: Google Play

¹Replication package: <https://github.com/mooselab/Sports-Apps-Analysis>, Password for Dataset: Sports123!

²<https://www.appbrain.com/>

³<https://www.appbrain.com/apps/popular/sports/>

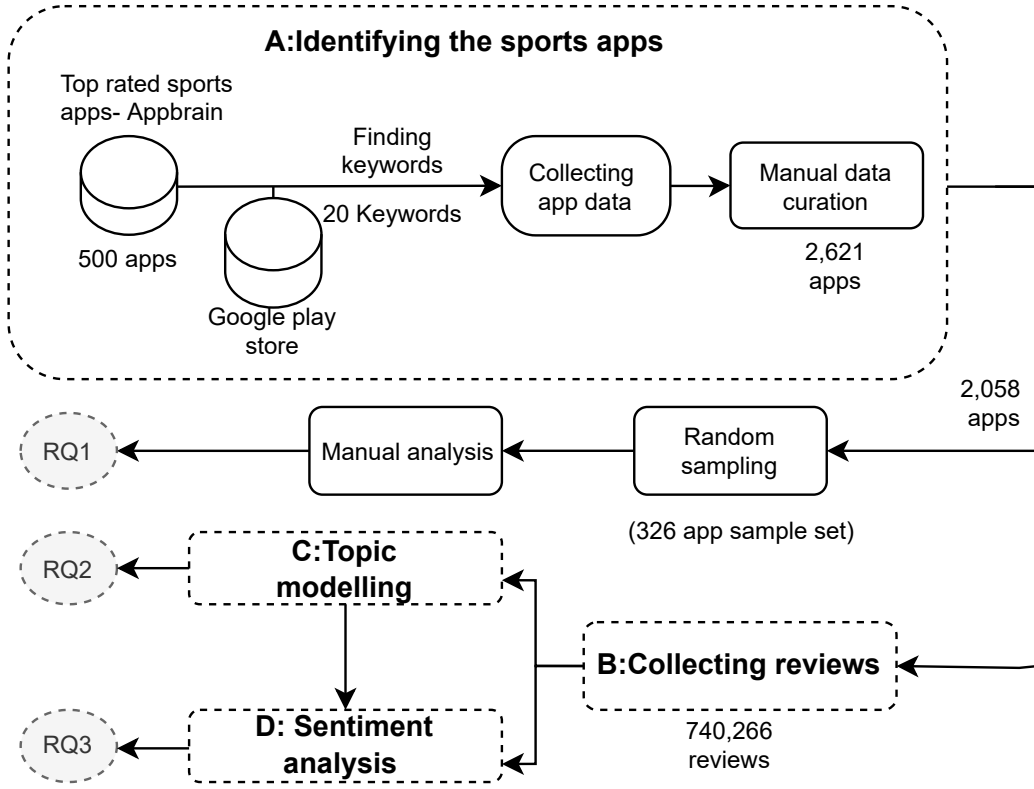


Figure 4.1 Overview of our study

URL, Short description, Launch time, Last update time, Rating, and Comment count. We used this set of *sports* apps as our reference set, and we used the Google Play URL to get a detailed description of these apps for our next step.

Finding keywords for sports apps

As using the keyword “sports” may retrieve many false positives (e.g., sports games) and the Google Play store limits the number of apps retrieved per request to 250, we need to identify a set of keywords which are used to search for *sports* apps in the store. We looked into the title and description of the top-rated sports apps and applied the keywords extraction algorithm YAKE [75] with a window size of 3. YAKE is an unsupervised keyword extraction method which provides a score for the contextually important words in a corpus. We select the most important word phrases, which can contain a maximum of 3 words. Each of the extracted keywords underwent a manual analysis to affirm its relevance to *sports* apps. We performed an iterative process to identify the keywords: the titles and descriptions of the apps searched from the identified keywords are further used to extract new keywords. We stopped the iterative process until no new sports-related keywords could be identified.

The step resulted in a final list of 20 sports-related keywords, including *golf GPS*, *sports betting*, *sports betting tips*, *soccer betting*, *fitivity*, *fitness*, *football*, *betting tips*, *soccer training*, *sports leagues*, *football tips*, *NBA League Pass*, *live sports*, *sports network Live*, *Pro sports subscription*, *state football*, *football training*, *champions league*, *premier league*, *sports player*.

Collecting app data

Using the keywords identified from the last step, we obtained 2,621 apps by searching in the Google Play store after removing duplication. Then, we extracted the Metadata of each app using Selenium HTTP requests. We focused on the following Metadata: Google Play URL, App name, Providers, Tags, Ads status, Description, Rating score, Rating Count, Downloads, Size, Date of update, Content rating permission, and Cost. The App name, Description, Rating score and Downloads are used in our RQ1 to study the characteristics of the sports apps. The Google Play URL is used to extract the app reviews that are used to answer our RQ2 and RQ3.

Manual data curation

Although we collected the 2,621 apps using sports-related keywords, some of the resulting apps may not be *sports* apps. For example, some *sports game* apps were included in the search results. Thus, we performed a manual verification of each of the resulting apps to remove false positives. In this manual process, we excluded all game apps and apps that are not related to sports, such as general streaming apps (e.g., Spotify). Our final verified dataset contains 2,058 *sports* apps.

4.2.2 Manual analysis

From our final set of 2,058 apps, we randomly sampled a set of 326 apps, which represent a confidence level of 95% and a confidence interval of 5%. The apps in the random set are manually examined to understand the characteristics of *sports* apps (i.e., RQ1), including their sports domains, main functionalities, and their used analytical methods. The details of our manual analysis process are described in Section 4.3.1.

4.2.3 Collecting reviews

We used the google-play-scraper API⁴ to extract the reviews of our final set of 2,058 *sports* apps. The API returned the following data we considered in this study: reviews and rating

⁴<https://pypi.org/project/google-play-scraper/>

Table 4.1 Cumulative number of reviews

Percentile Distribution	0	25	50	75	100
Number of Reviews	0	7	42	322	204,970

scores associated with them. In total, we extracted 740,226 reviews for all the apps. Table 4.1 shows five-number summary statistics and the average number of reviews of all apps that we study.

4.2.4 Topic Modeling

To understand the topics raised by sports app users in their reviews, we perform topic modeling to extract semantic topics from all the collected reviews. Then, we analyze the resulting topics to answer our RQ2. The details of our topic modeling and analysis approaches are described in Section 4.3.2.

4.2.5 Sentiment analysis of user reviews

We perform sentiment analysis of all the reviews of the collected *sports* apps to understand user-perceived negative opinions regarding the *sports* apps (i.e., RQ3). In particular, we analyze users’ sentiment in the reviews associated with each topic derived from our topic modeling process. The details of our sentiment analysis approach are described in Section 4.3.3.

4.3 Experiment Results

4.3.1 RQ1: What are the characteristics of sports apps?

Motivation

It is reported that there are over 46K apps in the *sports* category in the Google Play Store [1]. However, the characteristics of these *sports* apps (e.g., their provided functionalities) are not clear. Therefore, we first investigate the characteristics of these apps, which can provide us with better context when analyzing these apps in terms of their review topics in RQ2 (Section 4.3.2) and user-complained aspects RQ3 (Section 4.3.3). In particular, in this RQ, we investigate the types of these apps, their main functionalities and the analytical algorithms used in them. Our results can help researchers and practitioners better understand the domain of *sports* apps and provide context for our further analysis in RQ2 and RQ3.

Approach

In this RQ, we first study the store characteristics of *sports* apps, including their over-trend and popularity (in terms of downloads and ratings). Then we manually examine their intrinsic characteristics (non-store characteristics), including their targeted sport types, their main functionalities, and their used analytical algorithm types.

Overall trend of *sports* apps. We study the overall trend of sports apps to understand the growth of the sports app industry. We looked into the launch time of the 500 top-rated apps from Appbrain and counted the number of these apps created each year. We did not consider all the 2,058 apps because the Google Play Store does not provide the launch time of the apps. Thus, we limited our analysis to the top 500 apps from Appbrain, which provides the launch time. The launch time of these 500 apps ranges from 2010 to 2020 (the time when the data was collected).

Downloads and ratings of *sports* apps. We study the number of downloads and the ratings of the sports apps to understand their popularity and user-perceived quality. We considered all the 2,058 manually verified sports apps and extracted their download and rating information from the Google Play Store (see 4.2.1). In the Google Play Store, the average rating of an app is provided as a real number ranging from 1 to 5 (e.g., 4.5). The number of downloads is provided as a range value (e.g., 100K+ or 1M+). We analyze the distributions of the number of downloads and ratings of the apps in this RQ.

Categorization of *sports* apps. To understand the characteristics of sports apps and provide context for our further analysis, we manually examined a statistically representative sample from the 2,058 apps to derive a categorization of these apps from three perspectives:

- **Type of sport.** The target sport of the app, such as *football* or *golf*. An app is labelled as *general* if it does not target a specific type of sport. In case an app does not associate with any kind of sports but a generic purpose e.g., news app we classify it to *NA*.
- **Functionality.** The main functionality of the apps, such as *training* or *betting*. However, in An app with multiple functionalities, we only consider the primary functionality.
- **Type of analytics.** We observe that many sports apps use statistical or predictive analyses to achieve their goals (e.g., predicting match outcomes). Thus we categorize the type of analytics used in the apps. We use an “unknown” label when we could derive the type of analytics used in an app. Similarly we used “NA” if the application does not use any kind of analytical algorithms used in them.

We created a sample set of 336 apps with a confidence level of 95% and a confidence interval

of 5%. We collated the Application name and the description associated with it as the information needed to be examined. If the data is insufficient, we direct to the play store links to verify more details. The manual labelling was done in 3 stages by researchers with a software engineering background.

- First round, three researchers labelled 10% of apps together until we reached a common understanding.
- Secondly, the first author labelled 70% and then discussed with other researchers and reached a consensus.
- Finally, two researchers labelled 30% independently to measure the reliability. A consensus was reached for each label.

We used Cohen’s kappa score [76] to measure the inter-rater reliability. For the categorization of the sports type and the functionality, we achieved Cohen’s kappa score of 1, which indicates perfect reliability [77]. For the type of analytics, we achieved a Cohen’s kappa score of 0.81, which indicate a **strong** reliability [77].

Results

Sports apps have been increasing steadily in recent years. Figure 4.2 shows the number of apps in Appbrain’s top-500 list that were launched each year. The figure shows that the number of apps launched each year is continuously increasing from 2010 to 2020, with some minor fluctuations in the years 2015 and 2017. The trend shows that the apps launched per year improved almost 4 times from 2010 to 2020. Although the results only indicate the trend of the apps in Appbrain’s top-500 list, we can infer that the overall number of sports apps may follow a similar trend. Interestingly, the trend approximately matches the prediction that the sports technology industry is going to have an average annual growth of 16.8% ($(1 + 16.8)^{10} = 4.73$) [70]. It is reported that there are over 46K sports apps at the time of writing this paper [1]. The fast-growing trend motivates us to focus our studies on sports apps.

Sports apps have a large number of user groups. Users are generally positive about the quality of the sports apps, with the average rating of the sports apps higher than general apps in the Google Play Store. Table 4.2 shows the distribution of the number of downloads of the sports apps. 47.28% of the studied apps have more than 100K downloads, while only 13.08% of the studied apps have fewer than 1000 downloads. Prior work emphasizes the importance of user ratings of mobile apps, reporting that 69%

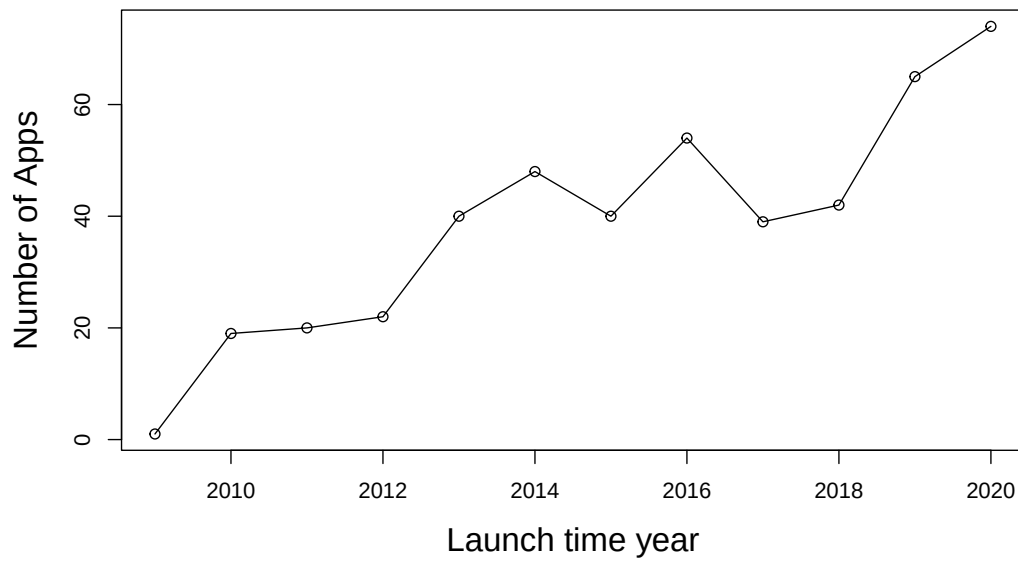


Figure 4.2 The number of sports apps launched per year in Appbrain's top-500 list.

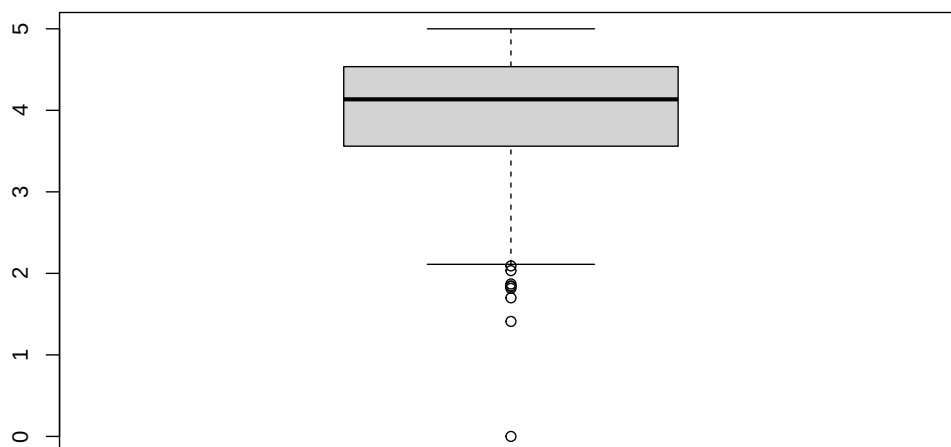


Figure 4.3 The distribution of rating scores of the studied sports apps.

of users consider app rating is very important [73]. Figure 4.3 shows the distribution of the average ratings of the studied sports apps. Out of the 2,058 sports apps that we evaluated,

Table 4.2 The distribution of the number of downloads of the studied sports apps.

Number of downloads	1-1000	1000-10k	10k-100k	100k-1M	1M+
Percentage of apps	13.08%	30.79%	8.85%	34.63%	12.65%

40% of them are rated above 4.5, while only 0.42% of them have ratings of less than 2. The average and median ratings of these apps are 4.0 and 4.1, respectively. The result indicates that users are generally positive towards the majority of these sports apps. Prior work on general sports apps reported that the average rating given for apps is 3.90 [66]. Thus sports apps have higher average ratings than general apps.

Table 4.3 shows our categorization of the sports apps in terms of the type of sport, the functionality, and the type of analytics.

Sports apps cover a wide spectrum of sports, with football⁵ being the most frequently targeted sport type. However, the distribution of sports apps is disproportional to the popularity of sports. From Table 4.3, we can understand that football (36.16%), fitness (15.25%), golf (9.04%), and basketball (5.08%) are the most common types of sports targeted by the sports apps. Besides, 23.73% of apps do not target a specific type of sport (i.e., the *general* category). According to a recent survey of the top fan-based sports by Sports Browser⁶, the sports with the most fan bases are in the order of football, cricket, basketball, field hockey, tennis, volleyball, table tennis, baseball, American football, and golf. 6/10 of these most popular sports are reflected in our studied apps in the sports app market while the other four popular sports (field hockey, tennis, volleyball, table tennis) are not. In addition, the number of apps for each sport is not proportional to the popularity of a sport. For example, while golf is the least popular in the top 10 popular sports, it has the second largest number of apps among the top 10 popular sports. The reason might be that users of golf apps are more willing to invest in the apps. Our results indicate the disproportional distribution of sports apps against the popularity of the sports.

Betting tips, training, and live updates are the most popular functionalities provided by sports apps. The functionalities derived from our manual analysis are defined below:

- **Betting Tips:** This functionality provides users with tips, suggestions and predictions to make bets on a future or ongoing sports competition. For example, *BetsWall Football*

⁵Football refers to association football or soccer in this paper.

⁶<https://sportsbrowser.net/most-popular-sports/>

Table 4.3 Characteristics of the manual analyzed sports apps: type of sports, main functionality, analytics approach. The results are obtained from analyzing a sample of 326 Sports apps.

<i>Classification category</i>	Type of sports							
Labels	Football	General	Fitness	Golf	Basketball	Car Racing	Martial Arts	American Football
Distribution	36.16%	23.73%	15.25%	9.04%	5.08%	1.69%	1.69%	1.13%
Labels	Baseball	Cricket	Athletics	Football and Basketball	Hockey	Wrestling	NA	
Distribution	1.13%	1.13%	0.56%	0.56%	0.56%	0.56%	1.13%	
<i>Classification category</i>	Functionality							
Labels	Betting Tips	Training	Live updates	Streaming	Tools	Tracking	Betting	
Distribution	24.86%	24.86%	15.25%	10.73%	7.34%	4.52%	3.39%	
Labels	News	Social Network	League Management	Radio	Team management	Health Tips		
Distribution	2.82%	2.26%	1.13%	1.13%	1.13%	0.56%		
<i>Classification category</i>	Analytical /statistical algorithm used							
Labels	Predictive analysis	Analytical statistics	Descriptive statistics	Analytical statistics and Predictive analysis	Unknown	NA		
Distribution	18.08%	7.34%	7.34%	3.95%	1.13%	62.15%		

The results are obtained from a manual analysis of 326 Sports by 3 researchers, obtaining an agreement (Cohen's Kappa) of 1 for Type of sports and Functionality. 0.81 for Analytical /statistical algorithm used.

Betting Tips provides football, basketball and tennis betting tips based on predictions made by artificial intelligence algorithms.

- **Training:** This functionality provide physical fitness training for different kinds of sports. For example, *Nike Training Club - Home workouts & fitness plans* provides home workouts or mindfulness training for the body and mind preparing for sports.
- **Live Updates:** This functionality provides live updates of the user-selected sports games as push notifications. For example, *BeSoccer - Soccer Live Score* covers more than 10,000 competitions and enables users to follow games live, and receive notifications of goals, line-ups, news, and signings.
- **Streaming:** This allows the users to watch sports games as a video stream. For example, *MyTeams by NBC Sports* provides easy access and availability of video sports information to the fans.
- **Tools:** This functionality makes use of the sensors in the phone and wearable to help sports players to measure their activities or functions to support sports. For example, the *YamaTrack Mobile* app provides accurate distances from tee to green and all points in between in a golf course.

- **Tracking:** Tracking provides end-to-end data collection and suggestions for sports players to monitor and analyze their activities. For example, *myCloudfitness* tracks all the fitness improvements and gives the necessary suggestions.
- **Betting:** This functionality enables users to place bets on competitive sports games. For example, *SuperDraft - a general book Free to Play for Prizes* allows users to place bets on sports events, including Pro and College Football, Pro and College Basketball, Pro and College Baseball, Hockey, MMA and Boxing, Soccer, Tennis.
- **News:** This functionality provides sports updates and news collected by news reporters. For example, *Onefootball - Soccer News, Scores & Stats* gives soccer news, scores, and stats of major competitions around the world, including the Premier League, La Liga, Serie A, etc.
- **Social Network:** This functionality allows different people interested in the same sports to connect and communicate together. For example, *TennisPAL* provides a platform that helps fans and followers of tennis all around the world to connect and share ideas about the sport.
- **League Management:** This functionality helps the users organize their own leagues or keep track and updates of their favourite leagues. For example, *Nizampur Premier League* helps Track all the local cricket leagues and manages the events in Nizampur.
- **Radio:** This functionality allows the users to get information about their favourite sports events in the form of audio streams. For example, *Pro Baseball Radio* delivers up-to-date scores, schedules, and standings, and allows users to listen to local radio stations for games live.
- **Team Management:** This functionality allows the managers of the sports to organize and track their team's activities in a convenient way. For example, *Soccer Tactics Board* helps the managers to track the progress and tactics of the team.
- **Health Tips:** This functionality helps the users get specific tips for health and injuries related to specific sports. For example, *Health Diet Foods fitness Help* provides a medical Guide (e.g., Health Diet Tips) specific to fitness programs.

Table 4.3 shows the distribution of the apps based on their main functionalities. Overall, the apps provide a number of similar functionalities. In particular, betting tips (24.86%), training (24.86%), and live updates (15.25%) are the major functionalities of the apps that we analyzed. The results indicate the high competitiveness in these areas and the importance

of achieving high quality for an app to succeed, which motivates us to analyze the topics that users are concerned about in their reviews (RQ2) and the aspects that users complain about (RQ3).

A significant portion of sports apps leverage statistical or predictive analyses.

The type of analytics derived from our manual analysis are defined below:

- **Predictive analysis:** Some apps use predictive models to make sports-related predictions. For example, *AI football betting tips* is a football betting tips & stats project which uses an artificial intelligence-based algorithm. The algorithm calculates all the probabilities and offers betting tips based on past data, team form, head-to-head results, goals scored, goals conceded, odds fluctuation, team profile (attacking or defending), and past playing formulas.
- **Analytical statistics:** Some apps analyze the data collected and help users interpret the results using predictions for a large group based on a representative sample of the group. For example, *mScorecard - Golf Scorecard* uses statistical analysis, which instantly calculates scores, handicaps, Stableford points and advanced round statistics.
- **Descriptive statistics:** Some apps use the data collected to summarise and visualize the data to understand the specific set of observations in them. For example, *Peak fitness* uses training plans and progress statistics to manage workout calendars, check clients in for workouts, and track current workouts.
- **Analytical statistics and Predictive analysis:** Some apps use both analytical statistics and predictive analysis. For example, *Bullet Bet predictions* uses machine learning algorithms (predictive analysis) combined with statistical analysis for more than 200 leagues. It uses machine learning algorithms which analyze historical results, home/away performance and team form to predict the winning team, as well as analytical statistics to understand the distribution against the hypothesis on the larger dataset.
- **Unknown:** Some apps do not mention the algorithms used but mention that they analyze data. For example, *Free Basketball betting Tips* says it provides Tips but did not mention the strategies for deriving them.
- **NA:** Some apps do not process data of any form.

As shown in Table 4.3, we observe that 36.72% of the analyzed apps use analytical methods. Sports analytics has gained fast-growing attention in recent years [78]. In our analysis, we

observed a significant portion of sports apps (e.g., betting apps) using statistical or predictive analysis to analyze the sports-related data collected, which suggests an interesting research area at the intersection of software engineering and data science in the context of sports apps: integrating data analytics in app development in a responsible way [79].

Sports apps have been increasing steadily in recent years, covering a broad spectrum of sports (e.g., football or fitness). In contrast, the distribution of sports apps is disproportional to the popularity of sports. We also observe that the sports app market is highly competitive, with many apps covering similar functionalities. Finally, we observe that a significant portion of sports apps uses statistical or predictive analytics, suggesting an interesting research area at the intersection of software engineering and data science in the context of sports apps.

4.3.2 RQ2: What are the topics raised by sports apps users in their reviews?

Motivation

Previous studies emphasized that understanding user reviews are important, and they contain a lot of information regarding the application behaviour and user impact [7, 30, 34, 35, 80]. These studies usually analyze the topics of general apps in the market. In comparison, this RQ aims to examine the user review topics of a specific app category (*sports* apps). In particular, we want to examine if focusing on such a specific category can provide different perspectives from a general analysis. A study specific to sports industry reviews and their impact has never been analyzed as per our knowledge.

Approach

Our studied 2,058 sports apps have a total of 740,226 reviews. It is difficult to manually examine all these reviews. Thus, we use automatic topic modeling to extract the topics from these reviews.

Topic modeling. To understand the topics the users are discussing, we use the *gensim* *mallet* implementation of the LDA topic modeling algorithm⁷, which has been used in many topic modeling-related research [45, 81, 82]. We performed data cleaning and preprocessing steps as follows. We removed special characters and emojis from the reviews. The reviews we considered are English text only. We created bigram and trigram models using the *gensim* model, followed by the stop words' removal and lemmatization. The *spacy* model was used

⁷<https://scikit-learn.org/sklearn.decomposition.LatentDirichletAllocation.html>

for lemmatization. After preprocessing, a set of words along with their frequencies in each review were produced and used in the topic modeling. We used the `mallet` 2.0.8 version.

Selection of the number of topics and other parameters. In topic modeling, selecting the number of topics is a critical task. We used the coherence score to select the number of topics (K). Specifically, we ran our topic modeling using K values from 2 to 50 with a step size of 1 and chose the K value with the highest coherence score. Out of all the 49 models, we selected the model with a K value of 17, which achieved the highest coherent score of 0.63. Besides, other parameters we used in the topic modeling include the optimization interval of ten, the number of iterations of 1000, the alpha value of 5.0, and the beta value of 0.01⁸.

Analysis of topics. Once we had the final topics, we manually labeled each topic. To label each topic, we looked into the representative words and the reviews that contributed most to that topic. During the manual labeling of topics, we merged topics that are highly similar to each other [83]. In particular, we are interested in how the resulting topics are different from the topics derived from the reviews of general apps. We also examined how the topics are distributed across different categories of sports apps in terms of the main functionalities. The labels defined were reviewed by four mobile app developers from a Google developer group (GDG Cochin). To validate the grouped topics for each review, we considered the topic percentage contribution metric. For each review, they have their topic contribution metric for the classified topic indicating the likelihood that a topic is represented in the text. We capped a threshold of 0.075 from the graph distribution of topic percentage contribution against the number of reviews. Value 0.075 covers 80% of the total number of reviews. The topics with the given topic percentage contribution or above are considered to be valid topics for the classified review.

Results

We derived 14 topics from the reviews of the sports apps. Our derived topics, along with their most significant words and the assigned labels, are listed in Table 4.4. *Positive feedback* is the dominant topic (19.27% in total reviews and 32.50% in the sample set) of the user reviews and it is the most common topic. In addition, *Up-to-dateness* (10.15%) and *Quality of Content* (10.67%) are among the frequently discussed topics, whereas *Bug Reports* (3.91%), *Negative Feedback* (3.85%) and *Advertisement* (3.45%) are the least discussed topics, as shown in Table 4.4. In the following section, we discuss each of the 14 topics and why they are significant for users and developers. In Table 4.4, we have added cross-references to the previous works where a similar topic has been discovered in previous works on app

⁸<https://senderle.github.io/topic-modeling-tool/documentation/2018/09/27/optional-settings.html>

review analysis.

1) **Positive feedback:** It is the most popular topic in user reviews. The users express their general positive feedback to an app in such reviews. Below is an example:

■ *"I love this app very much."*

2) **Quality of content:** Many sports apps deliver sports-related content (e.g., instructions, videos, news, or radio broadcasting) to users. The quality of such content is of critical importance to the users.

■ *"Even if you barely have time to work out this app will work for you Easy to understand, clean and clear instructional videos..."*

3) **Up-to-dateness:** Many types of sports apps need to provide regular updates to the delivered content to the users, such as apps that provide betting tips, live updates, or news. The timeliness of such updates is one of the most concerning topics for the users of these apps. **This is a new review topic that has not been previously identified in other studies of app reviews.**

■ *"This app helps me get notifications of scores and when matches are starting because if I don't have WiFi, they will just send me I'd somebody scores or I'd a match started."*

4) **Return on investment:** 24% of the sports apps in our dataset are paid and developed as a business. Even when an app is free, the users usually need to invest time in the advertisement. The users are concerned about the value or performance of the app based on the money or time they invest in it.

■ *"It's designed to take as much of your hard-earned cash as possible. Constantly need energy drinks, and your salary just doesn't keep up with the cost if you pay for the full ad-free version, then this shouldn't be as bad as if playing for free."*

5) **Version update:** This topic mainly discusses the challenges the users face with version updates or the differences between the two versions, compatibility of the version with devices and use case. And this will help us, the developers, to decide the impacts of each version and mitigate and resolve issues associated with them.

Table 4.4 Topics extracted from the reviews of sports apps.

Topic Number	Topic Label	Key Words	Total Con.	Prior work
1	Positive feedback*	great, easy, love, helpful, follow, app, simple, fun, super, fast good, app, nice, world, job, graphic, usefull, enjoy, awesome, pretty app, good, great, awesome, excellent, cool, experience, enjoy, site, woman	19.67%	[49], [44]
2	Quality of content*	love, amazing, perfect, fitness, recommend, enjoy, daily, yoga, absolutely, class workout, exercise, plan, home, gym, set, training, routine, choose, rest	10.67%	[44]
3	Up-to-dateness	game, team, match, play, score, football, live, news, player, fan	10.15%	New
4	Return on investment	free, pay, money, subscription, month, version, year, worth, buy, cancel	8.86%	[3],[33] [84], [44]
5	Version update	work, update, phone, android, screen, fine, version, late, slow, device	7.62%	[66],[33]
6	Streaming quality	watch, tv, show, live, stream, channel, service, video, minute, freeze	6.91%	[66]
7	Impact of application	day, feel, lose, week, weight, start, body, result, exercise, fit	6.54%	[44]
8	Feature request	option, video, notification, user, put, view, find, add, read, turn	5.37%	[85], [3],[49] [33]
9	Accuracy of prediction	application, tip, bet, prediction, guy, win, amazing, good, accurate, vip	5.27%	New
10	Review credibility	app, give, download, star, sport, information, review, rate, full, info	4.30%	[7]
11	Bug reports	fix, problem, issue, open, log, load, crash, sign, account, email	3.91%	[85], [49], [84]
12	Negative feedback	make, worst, app, find, change, thing, issues, program, people, hate	3.85%	[6]
13	Advertisements	time, ad, bad, back, start, long, waste, close, stop, annoying	3.45%	[3], [84]
14	Precision of tools	add, track, feature, run, show, datum, step, point, save, record	3.43%	New

*The topic is merged from multiple LDA-produced topics in our manual labelling. Therefore, there are multiple sets of keywords. The topics have a Cohen kappa of 1 and are validated by developers from the industry.

■ “update won’t install on note 9”

6) **Streaming quality:** Watching sports videos live is vital for sports fans. The quality of the streaming bothers them and thus becomes the topic in their reviews. Audio broadcasting

through radio to discuss sports reviews also contributes to the given topic.

"The app needs work. This review is for the app and not the fight pass service. There is no option for offline viewing, so every video is a stream, which this app sucks at. You can't fast-forward without the playback freezing and having to start over. Live streams that I've watched were VERY problematic, and important exchanges were missed due to freezing. If I'm expected to pay this much for a service this frustrating app can't deliver, what's the point? "

7) **Impact of application:** Many sports apps (e.g., training or tracking apps) can have certain physical or mental impacts (e.g., improving fitness) on the users. The impacts of such apps are an important topic raised in users' reviews.

"Lose Weight and Burn Belly Fat with fat burning workouts for women at home. The best lose weight app for women to burn fat and lose weight at home! With simple and effective fat-burning workouts for women, you can lose belly fat, lose thigh and arm fat. Follow the 30-day plan and take just a few minutes a day to lose weight and get in better shape! 2-7 min fast workouts and HIIT workouts allow you to lose weight and keep fit anytime, anywhere. No excuse anymore! You can track burned calories"

8) **Feature request:** Users not only express opinions on the existing features but also tend to convey what they like through their reviews. User reviews on this topic itself will give companies a plethora of ideas to improve their apps. Prior studies [3, 49, 85] have proposed approaches to mine feature requests from user reviews.

"Awesome app for beginner, But need some offline video feature otherwise this is great."

9) **Accuracy of prediction:** From our analysis in RQ1, we understood that over 22% of apps (e.g., betting tip apps) use some form of predictive analysis. For those apps, the accuracy of the predictive analysis is an important topic raised by users in their reviews. **This is a new review topic identified in our work.** Although the topic name "accuracy" has been mentioned in works like [6], however, the context of using accuracy in our work is mostly on the accuracy of the predictions in the concerned apps.

"Honestly, a correct prediction is almost impossible bcz you can't control what goes on in the field of play, but this app has been amazing..I give u 5 stars. good job guys!!"

10) **Review credibility:** In the reviews on this topic, the users validate the previous reviews with their own personal experiences.

"Can't trust reviews. All of the reviews are five stars and tons of them are all posted on the same day as if it's the employees or paid reviews. Just suspicious. I downloaded it and immediately can't even see if I want the app before it asks for you to pay for a year. I then really check out the reviews and they look fake. After every semi-negative review, they flood the app with tons of positive ones. So SUS"

11) **Bug reports:** While testing apps, there are always constraints about testing all the scenarios. But the users are the best testers out there with a different setup. They will explain the challenges to using the apps, and those can be resolved for the better performance of the apps.

"When system errors occur, app doesn't save your place. The user has to start over and if it's been more than a few minutes into the workout, the app finishes the workout, which is more hassle to redo it. "

12) **Negative feedback:** In contrary to the "positive feedback" topic, users express their general negative feedback to an app in reviews on this topic. They typically use short messages to express their frustration like the one below.

"Most disappointing app I have used in my life"

13) **Advertisement:** From our previous analysis, we understood that the majority of the apps are free, and the profit of the apps generally comes from the ads in the apps. However, the ads have a huge impact on the user experience which is discussed by the users in their reviews.

"Would have given it a 4star but come on, what is it with these "ads?" Ads keeps popping up even if I close it, the only way out is to "force stop" the app, delete data and run the app again. This happened to me last year until an update put everything back to normal, now it's started again, showing Guinness ads over and over again, I don't drink Guinness for God sake.."

14) **Precision of tools:** As discussed in RQ1, there are mobile apps which take leverage of the sensors in mobile phones or wearables to measure distance, speed etc. and act as a tool to support the sportsmen. The precision of the tools is vital for the success of the application and thus an important concern of the users as expressed in their reviews. As the sensors of different mobile devices may have different accuracy, it is important that app developers test their apps on mobile devices of diverse sensing capabilities.

“Tracks step count during a run but not cadence (steps/min) Spoken announcements only at 0.5km and 1km; no 400m option or ability to enter a custom distance graphs are tiny and coarse like Samsung Health but at least the plot points here are selectable. Missed the opportunity to match plot points with points on map like in Nike's running app No auto-pause, not even a simple one let alone a good one like Strava's No split or lap times All in all, very basic as a running app and can't stand alone.”

While most of these topics are found in previous review analysis studies, we identify new topics that are special for sports apps, including *accuracy of prediction*, *up-to-dateness*, and *precision of tools*. Prior works have analyzed app reviews to identify the topics in users' reviews. They either analyze apps across different app categories (e.g., [26, 30, 67]) or focus on a small set of particular apps (e.g., [68, 69]). In comparison, this work analyzes the review topics of a large number of apps in a single app category (*sports apps*). We observed that, out of the 14 topics derived from our analysis, 11 topics could be mapped to similar topics identified in previous works. For example, we recognized topics such as the impact of application which has been the focus of study on its own [68, 69]. However, our analysis of the review topics of such a single app category can provide more specific insights into the app category, as we identified three review topics (*accuracy of prediction*, *up-to-dateness*, and *precision of tools*) that are associated with the characteristics of the sports apps. For example, as many sports apps (e.g., betting tip apps, live update apps, etc.) need to provide regular updates to the delivered content, the up-to-dateness of the content of the apps is one of the most concerned topics of the users. Our results suggest that future work on app review analysis can pay more attention to analyzing a large number of apps in individual app categories.

The distributions of review topics vary across apps with different main functionalities. Fig 4.4 shows the analysis of the reviews over different functionalities and different topics. As we concluded previously, the *negative feedback* bubbles overall functionalities are relatively small, whereas *positive feedback* is bigger, thus the users are generally happy about sports apps and their use cases. The topic *accuracy of prediction* has a significantly higher distribution for apps with the main functionality of *betting tips* which indicates the importance of using more accurate machine learning for these apps. Similarly, topic *precision of tools* is highly cluttered in the functionality *tracking* which uses support/information from build sensors of phones or wearable devices, which indicates the necessity of adaptability of applications to the cutting-edge technologies of wearable devices or general IOTs. The topic *up-to-dateness* is mainly associated with two functionalities *news* and *live updates*, which deals with the timeliness delivery of the information to the users, which indicates the impor-

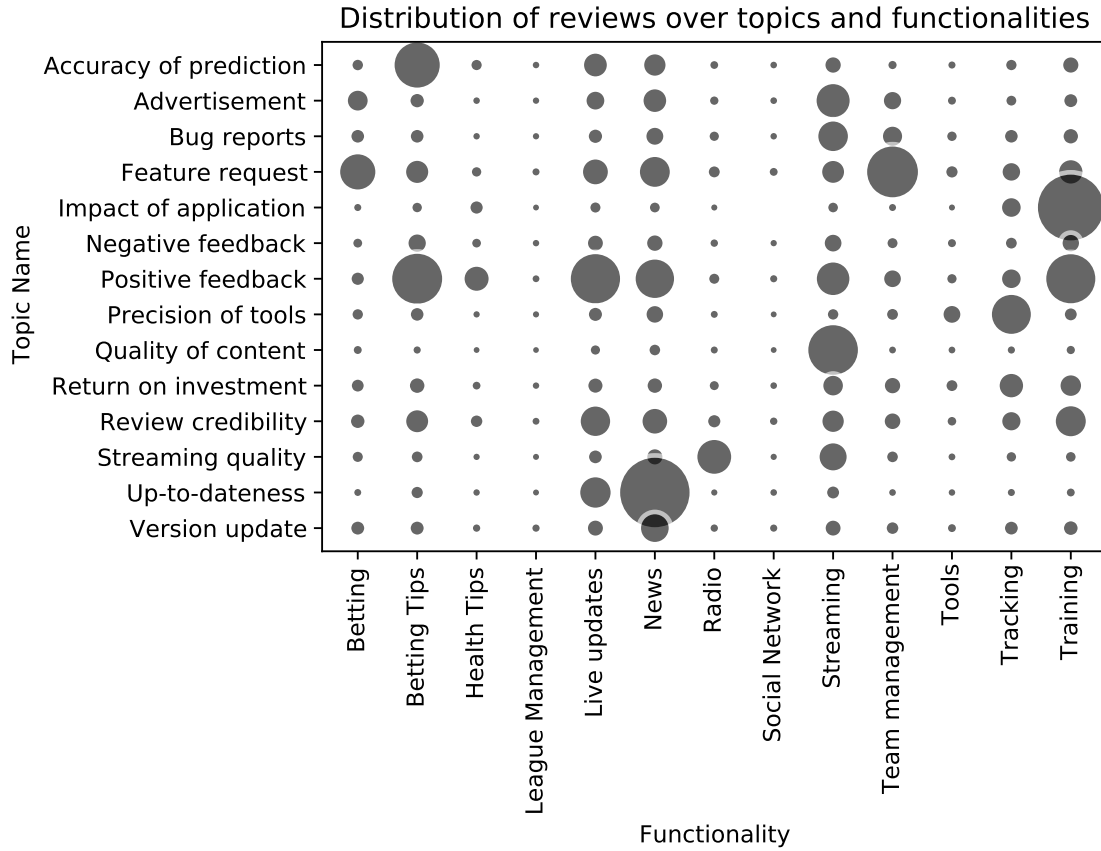


Figure 4.4 Distribution of reviews over different functionalities and topics.

tance of keeping the information updated and push notifications in these mobile applications.

We derived 14 topics from the user reviews of 2,058 sports apps. While most of these topics are found in previous review analysis studies; we identify new topics that are special for sports apps, including *accuracy of prediction*, *up-to-dateness*, and *precision of tools*. Our results suggest that future work on app review analysis can pay more attention to analyzing a large number of apps in individual app categories.

4.3.3 RQ3: What do users complain about sports apps?

Motivation

App users express their feedback about an app in their reviews. Understanding what users complain about can provide insights for app developers to improve their apps' quality or meet the users' needs [3, 6, 33]. In RQ2, we identified the topics that users of sports apps

discuss in their reviews. In this RQ, we further analyze which factors (topics) users complain about, accounting for both the general topics (e.g., quality of content) and the topics specific to sports apps (e.g., up-to-dateness).

Approach

In this RQ, we analyzed the existing user review analysis method to understand why users complain about sports apps. The previous study [6] has been done on the generic apps to understand what users complain about the apps [6]. Each review is associated with a rating, the quantifiable measure of the users' feedback about an app. The review is the descriptive way of the same feedback. Hypothetically, both are significantly correlated: the reviews with low ratings should be bad, and those with high ratings should be good. However, there is a lot of noise in Text-Rating-Consistency(TRC) [6]. The approach we used to mitigate this challenge is to evaluate the sentiment of the reviews and correlate it with the review ratings available.

The intuition is that negative sentiment is associated with negative feedback in the user review. Prior work has used a similar sentiment analysis of user reviews to understand how users like the features of an app [8].

Sentiment analysis. We use NLTK's pre-trained sentimental analysis library called VADER (Valence Aware Dictionary and sEntiment Reasoner)⁹. VADER appears to be the best suit for the language used in social media [86], which are short sentences with some slang and abbreviations expected to be similar in user reviews. The median review length of the entire review corpus under study is 12 words per review sentence. The results mapped between 1 and -1, where 1 indicates entirely positive, -1 indicates entirely negative, and 0 indicates neutral reviews. Following the guideline¹⁰ by VADER, we use the *compound* sentiment score to classify the sentiment: positive sentiment (score ≥ 0.05), neutral sentiment ($0.05 > \text{score} > -0.05$), and negative sentiment (score ≤ -0.05), similar to prior work [87]. Although different adaptive ranges of sentiments were selected for the reviews based on the distribution by [86, 88], we are sticking to the basic conventions. We also experimented with sentiment analysis tools such as the Textblob library¹¹. Our results show that VADER gives better results which confirm others' observations¹².

Correlation analysis. To understand the relationship between a user's sentiment in a

⁹<https://github.com/cjhutto/vaderSentiment>

¹⁰<https://github.com/cjhutto/vaderSentiment#about-the-scoring>

¹¹<https://textblob.readthedocs.io/en/dev/>

¹²<https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/>

review and the rating given to the reviews by the users, we perform a pairwise Spearman correlation analysis between the following metrics:

- **Average user rating:** Arithmetic mean of the ratings associated with all reviews of sports apps for each app.
- **Average sentiment:** Arithmetic mean of sentiment calculated for all reviews for each app.
- **Percentage negative rating:** Distribution of negative rating (≤ 2) to the number of all reviews associated with each app.
- **Percentage negative sentiment:** Distribution of negative rating (≤ -0.05) to the number of all reviews associated with each app.

For each pair of metrics, we calculate Spearman’s correlation coefficient and the p-value [89]. The results are shown in Fig. 4.5.

Removing outliers. User reviews are one of the pure forms of natural language; it contains a lot of noise, and the tricky part of natural language - is the irony. To get rid of the outlier, we used the sentiment score and the user rating of the user review to detect the inconsistency between them. If the rating is less than or equal to 2 (indicating a negative review [26, 33, 73, 90]) and the sentiment score of the review is not less than -0.05, we consider it as an outlier and remove it from our analysis.

Functionality dimension analysis. We further analyze the user-complained factors for each group of apps with the same main functionalities, we indicated it as functionality dimension analysis. Understanding the complained factors on a functionality dimension gave more intricate information about the concerns users have for a specific group of apps. Functionality dimension analysis is a comparison of the topics generated for each review to the functionality of the sports application with which the reviews are associated with. For each group of apps with the same main functionality (from the manually studied sample in RQ1, see Sec 4.2.2), we analyzed the average sentiment of the reviews and the percentage of negative sentiment reviews associated with each topic.

Result

Users are generally happy about the sports application and their performance. Figure 4.6 describes the distribution of reviews over the review sentiment values and the user ratings. Overall, 94.44% of the reviews have a positive sentiment (i.e., sentiment score $\geq$

0.05), while only 5.56% of the reviews have a negative sentiment (i.e., sentiment score ≤ -0.05). 65.5% of the reviews with negative sentiment are associated with negative ratings (e.g., rating ≤ 2). In comparison, only 34.5% of the reviews with positive sentiment are associated with negative ratings. As we can see, there are a few outlier instances with very negative sentiment reviews (e.g., “Not working”, with a sentiment of -0.9) but with a very high (e.g., a rating of 5). To mitigate the text rating inconsistency, we removed the outliers to make the analysis fairer, as described in the approach section.

Reviews with negative rating scores are highly correlated with negative sentiment reviews. In Fig 4.5, we analyzed Spearman’s rank correlation to understand the correlation between the metrics such as *Average user rating*, *Average sentiment*, *Percentage negative rating*, *Percentage negative sentiment*. The results show that user ratings are highly correlated with their sentiment in the corresponding reviews. For example, the percentage of negative sentiment reviews and the percentage of negative rating scores of the studied apps have a correlation of 0.99 ($p < 0.001$). Besides, the average sentiment of the reviews and the average rating scores of the studied apps have a correlation of 0.99 ($p < 0.001$).

Users are mostly complaining about the advertisements and the quality (bugs, content quality, or streaming quality) of sports apps. Table 4.5 shows the average sentiment of the reviews associated with each topic and the percentage of reviews with negative sentiment (i.e., score ≤ -0.05), as well as the average ratings and the percentage of reviews with a negative rating (rating ≤ 2) associated with each topic. As expected, the reviews associated with the topics of *advertisement*, *bug reports*, and *negative feedback* have the lowest average sentiment, which confirms the results of prior studies [3, 6, 33]. Among the other review topics, the ones with the lowest average sentiment values include the topics such as *quality of content*, *version update*, *return on investment*, and *streaming quality*. The topics of *quality of content* and *version updates* have been identified in prior studies as topics [3, 6, 33]; however, they are not identified as negative aspects in these studies. Our results indicate that these are specific to sports apps. The topics *return on investment* and *streaming quality* are also recognized for generic apps as negative sentiment topics [6].

The following are the review topics associated with the most negative sentiment:

Advertisement: 76% of the apps that we analyzed are free applications (completely or with premium purchases inside). Most of the revenue of these applications comes from advertisements. Our results show the negative implications of having advertisements on user experiences. Thus, the question of compromising user satisfaction for income needs to be considered by developers.

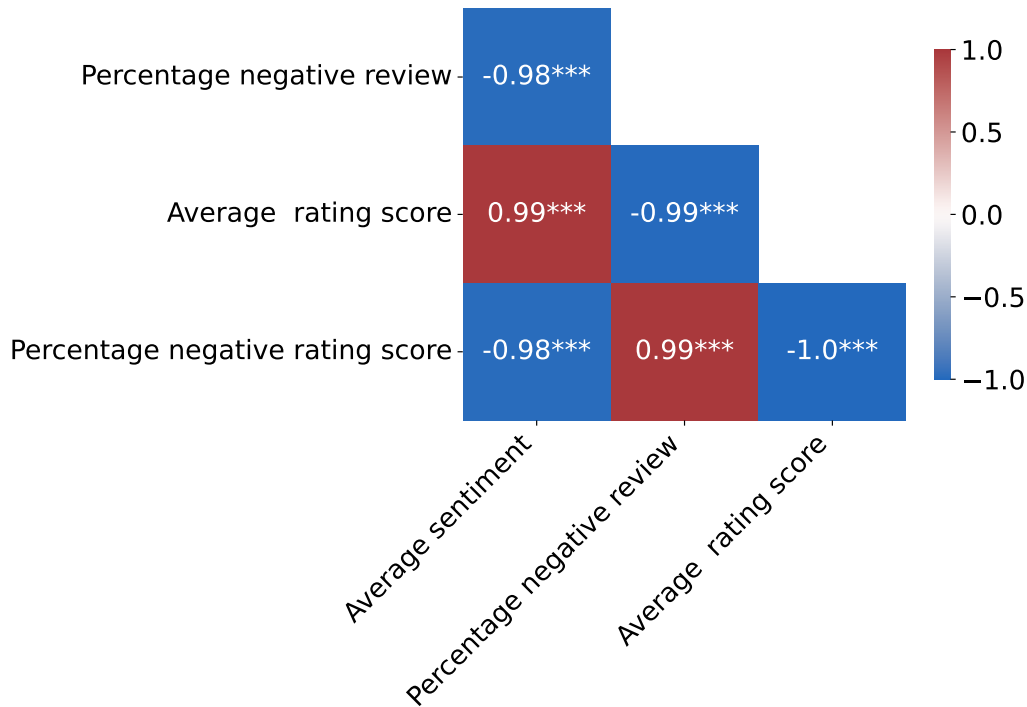


Figure 4.5 Pairwise Spearman correlation between Average sentiment, Percentage of negative reviews, Average rating score and Percentage of negative rating score of the reviews of each apps(after removing outliers). The * symbols along side the correlation values indicate the statistical significance of the respective correlation: No symbol $p \geq 0.05$, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Bug Reports: Irrespective of the challenges with natural language analysis, some users are very intricate about their issues and inform it via reviews. The sentiment associated with the reviews indicates that they are not happy about it(4.4) however, users provide detailed issues, including how to replicate and what is the expected behaviour– which helps the developers mitigate the issue faster.

Quality of Content: Table 4.4 shows that this is the second most popular topic among the reviews. From manually examining the reviews, it is clear that **both subjective (e.g., content appropriate to user groups) and objective (e.g., network metrics and HTTPS streaming metrics)** aspects are user concerns that developers have to put more focus on.

Negative Feedback: This is a prominent topic with an expected negative sentiment, which

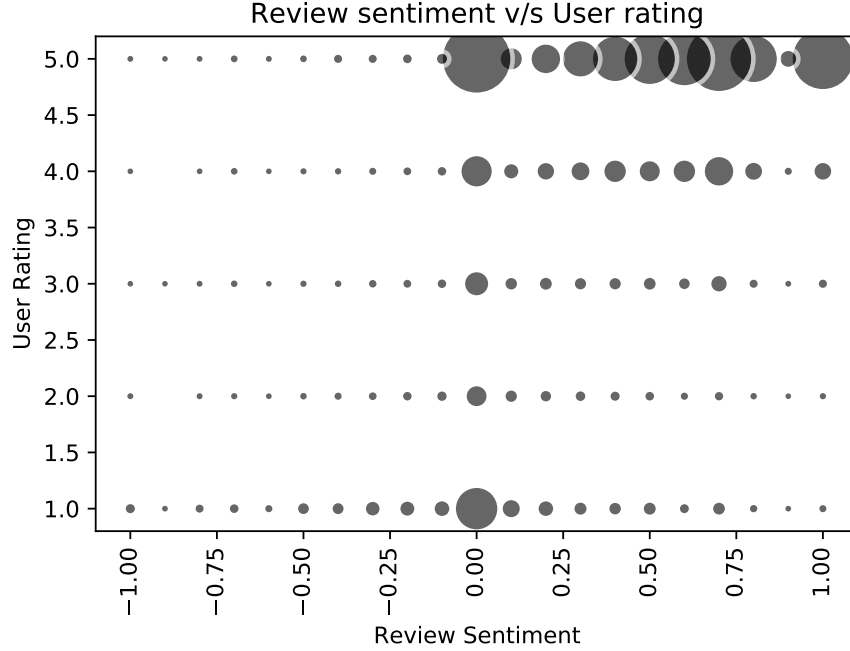


Figure 4.6 Distribution of review sentiment over user rating associated with reviews (before reviewing the outliers).

demonstrates the validity of our topic modeling results [91].

***Streaming, Betting, Team management and League management* are the most complained sports app functionalities.** Table 4.6 shows that the percentage of negative sentiment reviews is 32.12%, 21.23%, 18.18% and 17.65% for the apps associated with the functionalities of *streaming*, *Betting*, *Team management* and *League management*, respectively. The most complained topics of these apps include *advertisement*, *bug reports*, *quality of content*, and *negative feedback*, and *version update*. The results indicate that users are concerned about the advertisement and quality (e.g., bugs or quality of content) of these functionalities supported by the sports apps. Future efforts are needed to improve the quality of these apps or explore other ways of revenue other than the advertisement.

In general, the most complained topics are consistent across the sports apps with different functionalities. Table 4.6 shows the review sentiment for each review topic of each app functionality. The topics *advertisement*, *bug reports*, *negative feedback*, and *quality of content* are consistently among the most complained topics of each functionality. For example, these four topics are the most complained topics of the most popular functionalities (Table 4.3): *Betting Tips* and *Training* apps. However, there are some exceptions. For example, the topic of *streaming quality* is among the most complained topics of the *News* apps,

Table 4.5 Average sentiment and average rating score of reviews with distribution of negative sentiment reviews and negative rated reviews across different topics. Table is sorted based on the the distribution of negative sentiment review distribution across topics

Topics	Percentage negative sentiment review	Average sentiment	Percentage negative rating score	Average rating score
Advertisement	35.66%	-0.02	59.60%	2.4
Bug reports	33.67%	0.06	50.90%	2.71
Quality of content	30.48%	0.17	45.40%	2.94
Negative feedback	29.02%	0.16	35.86%	3.31
Version update	21.72%	0.23	26.33%	3.54
Return on investment	20.28%	0.31	27.87%	3.65
Streaming quality	19.78%	0.3	25.62%	3.78
Feature request	10.74%	0.42	14.81%	4.18
Precision of tools	7.35%	0.52	8.10%	4.37
Accuracy of prediction	7.27%	0.43	9.98%	4.29
Impact of application	4.42%	0.56	3.23%	4.63
Up-to-dateness	3.33%	0.52	4.32%	4.63
Review credibility	2.51%	0.6	3.68%	4.71
Positive feedback	2.51%	0.51	4.08%	4.68

which indicates users’ concerns regarding the quality of *News* apps’ multi-media streams.

Users are generally positive about sports apps, as indicated in their review sentiment. On the other hand, they are mainly complaining about the advertisements and quality (bugs, content quality, streaming quality) of these apps, which are consistent among the apps of different functionalities. In particular, *Streaming*, *Betting*, *Team management* and *League management* are the most complained about sports app functionalities.

4.4 Discussions

In this section, we discuss the implications of our results.

Sports apps are a rapid-growing and competitive market. Our results in RQ1 indicate that the number of sports apps are steadily increasing in recent years. It is reported that there are over 46K sports apps on the Google Play Store [1]. On the other hand, there are many sports apps targeting the same sports types or implementing similar functionalities. For example, only out of our sampled 336 apps, there are 40 apps and 22 apps that provide betting tips and live updates for football, respectively. The competitive nature of the sports app market suggests that developers need to pay extra attention to the user-perceived quality, which is communicated through their ratings and review text. Thus, in this work, we study

Table 4.6 Distribution of average review sentiment over different topics and different functionalities of sports apps. Also the percentage of negative sentiment reviews out of all reviews.

Topic name	Betting	Betting Tips	Health Tips	League Management	Live updates	News	Radio	Social Network	Streaming	Team management	Tools	Tracking	Training	Grand Total
Accuracy of prediction	0.38(13.1%)	0.49(5.71%)	0.46(2.53%)	0.54(0%)	0.38(4.48%)	0.43(6.12%)	0.21(12.12%)	0.4(14.29%)	0.06(22.39%)	0.29(20%)	0.33(4.55%)	0.17(7.41%)	0.41(9.64%)	0.43(7.27%)
Advertise-ment	-0.06(33.33%)	0.08(25.33%)	0.09(30.77%)	-0.07(40%)	0.06(24.91%)	0.08(27.05%)	-0.06(51.16%)	-0.17(53.33%)	-0.11(45.66%)	-0.12(46.51%)	0.04(16.67%)	0.17(21.13%)	0.25(20.86%)	-0.02(35.66%)
Bug reports	-0.03(36.84%)	0.18(20.3%)	0.39(0%)	0.07(33.33%)	0.24(14.19%)	0.18(26.02%)	0.24(21.67%)	0.39(8.33%)	-0.09(47.49%)	-0.03(37.54%)	0.13(28.57%)	0.25(21.97%)	0.22(22.16%)	0.06(33.67%)
Feature Request	0.3(19.51%)	0.43(3.8%)	0.49(1.61%)	0.34(14.29%)	0.42(4.17%)	0.47(8.79%)	0.54(6.38%)	0.62(8.57%)	0.22(17.48%)	0.43(12.11%)	0.44(5.94%)	0.61(3.03%)	0.55(3.85%)	0.42(10.74%)
Impact of Application	0.31(9.52%)	0.37(6.35%)	0.61(0.81%)	0.23(0%)	0.35(5.26%)	0.44(4.29%)	0.43(16.67%)	No reviews	0(28.36%)	0.4(10%)	0.27(20%)	0.59(4%)	0.58(3.95%)	0.56(4.42%)
Negative Feedback	-0.12(50%)	0.24(20.08%)	0.45(8%)	-0.28(54.55%)	0.16(20.63%)	0.28(24.14%)	0.05(48.15%)	0.02(33.33%)	-0.23(57.68%)	-0.13(55.41%)	0(37.84%)	0.35(16.48%)	0.44(11.64%)	0.16(29.02%)
Positive Feedback	0.53(2.44%)	0.53(0.77%)	0.51(0.2%)	0.34(11.76%)	0.52(0.62%)	0.6(1.98%)	0.57(1.33%)	0.52(4.55%)	0.22(17.63%)	0.53(1.25%)	0.5(1.67%)	0.61(1.34%)	0.55(0.57%)	0.51(2.51%)
Precision of Tools	0.27(19.51%)	0.35(3.03%)	0.45(11.11%)	0.49(0%)	0.36(6.29%)	0.42(0.75%)	0.42(15.79%)	0.47(0%)	0.16(25.3%)	0.45(8.7%)	0.43(10.7%)	0.65(4.83%)	0.47(5.22%)	0.52(7.35%)
Quality of Content	-0.02(27.78%)	0.21(4.35%)	0.22(0%)	-0.05(100%)	0.28(6.78%)	0.31(13.79%)	0.2(25%)	-0.37(50%)	0.16(32.71%)	-0.06(29.41%)	0.05(37.5%)	0.35(17.39%)	0.19(17.07%)	0.17(30.48%)
Return on Investment	0.18(22.81%)	0.32(18.89%)	0.41(6.06%)	0.52(18.18%)	0.22(15.29%)	0.37(14.04%)	0.32(25.45%)	0.38(28.57%)	0.1(32.83%)	0.11(35.82%)	0.28(28.26%)	0.48(14.01%)	0.45(11.85%)	0.31(20.28%)
Review credibility	0.62(1.96%)	0.6(1.93%)	0.65(1.89%)	0.61(5.56%)	0.56(2.33%)	0.64(1.73%)	0.66(1.64%)	0.63(0%)	0.44(9.41%)	0.64(0.96%)	0.59(4.26%)	0.6(2.09%)	0.65(0.79%)	0.6(2.51%)
Up-to-dateness	0.28(16.88%)	0.43(2.22%)	0.57(0%)	0.44(0%)	0.3(8.15%)	0.43(12.76%)	0.37(17.89%)	0.48(25%)	0.11(33.49%)	0.36(14.77%)	0.29(29.41%)	0.48(3.17%)	0.33(13.43%)	0.3(19.78%)
Version update	0.21(10.53%)	0.41(0.99%)	0.4(0%)	-0.08(50%)	0.46(2.58%)	0.54(3.19%)	0.66(0%)	0.35(0%)	0.19(11.21%)	0.37(13.33%)	0.29(8.33%)	0.41(0.52%)	0.43(0%)	0.52(3.33%)
Streaming quality	0.13(25.18%)	0.26(9.29%)	0.52(0%)	0.44(12.5%)	0.3(10.42%)	0.22(26.66%)	0.27(11.54%)	0.28(21.74%)	0.01(37.04%)	0.13(32.41%)	0.18(17.14%)	0.36(15.44%)	0.36(10.26%)	0.23(21.72%)
Grand Total	0.23(21.23%)	0.46(5.36%)	0.52(1.75%)	0.32(17.65%)	0.43(5.05%)	0.47(8.25%)	0.38(16.71%)	0.4(15.15%)	0.1(32.12%)	0.33(18.18%)	0.34(15.14%)	0.56(7.18%)	0.54(4.52%)	

the reviews and ratings of sports apps to understand users' review topics and the aspects that users complain about these apps. Our results can provide guidance for developers to improve the user-perceived quality of their apps and win the competition in the market (e.g., by reducing the negative impact of advertisements and improving the quality of content). We also encourage future work to study the factors that influence the competitiveness of sports apps in the competitive market.

Sports apps provide good examples of studying the integration of data science or machine learning techniques in software applications. As discussed in RQ1, a significant portion (37.85%) of sports apps leverage statistical or predictive analyses (*a.k.a.*, sports analytics). For example, sports apps providing betting tips, training, and tracking functionalities may rely on statistical or predictive analyses to help users make data-driven decisions. In RQ2, we also observe that users are concerned about the topic of the *accuracy of prediction*. Our findings highlight the significant applications of statistical and predictive analyses in sports apps, which suggest that sports apps may be explored in future work to study the engineering and integration of data science or machine learning techniques in real-world software applications.

Analyzing a targeted category of apps (e.g., sports apps) can provide more specific insights than analyzing apps across different categories while still being relevant for a large number of apps (e.g., tens of thousands of apps in a category). In this work, we started a large number (2,058) of apps in sports apps in contrast to traditional large-scale studies that study apps in the app store regardless of their categories. Such a study can provide more specific insights for the app category, for example, the characteristics of the sub-categories (e.g., by sports types or by the provided functionalities), the review topics (e.g., the review topics such as *accuracy of prediction*, *up-to-dateness*, *precision of tools* that are new to the sports apps), and the aspects that users complain about. Although the results may be limited to a specific app category, as an app category typically include thousands to hundreds of thousands of apps, the results can still be relevant to a larger number of apps and developers.

4.5 Threats to Validity

This section discusses threats to the validity of our study.

4.5.1 Internal Validity

In this work, we use automated topic modeling (LDA) to derive topics from user reviews based on the co-occurrence patterns of words in the reviews. LDA categories the words that frequently co-occur together into a same topic. This is because the probability of a word occurring in a document is influenced not only by the topics present in the document but also by the frequency of co-occurrence of that word with other words in the same document. To estimate the topic distributions and associated word distributions, LDA uses a Bayesian inference algorithm that iteratively updates the topic assignments of words in each document and estimates the topic distributions and word distributions that best explain the observed data. We use a coherence score to select the optimal number of topics. The automatically derived topics may not represent the exact topics in the user reviews. To mitigate the threat, we manually analyzed the representative reviews of each topic to examine the validity of the topic modeling results and merge some similar topics based on prior practices[83].

4.5.2 External Validity

This work studies 2,058 sports apps which is a small subset of all the sports apps (over 46K according to AppBrain [1]). The scope of our study is limited by the fact Google Play limits access to the complete list of apps in an app category. To maximize the number of sports apps studied in this work, we started from the top 500 popular sports apps provided by AppBrain and used an iterative process to identify new keywords and search for additional sports apps. We also manually verified each of the resulting apps to filter out the ones that do not belong to the category. Our obtained apps cover a variety of sports categories and app functionalities.

4.5.3 Construct Validity

In RQ1, we manually studied a sample of sports apps to understand their characteristics. The results may be biased by the individuals who performed the analysis. To mitigate the bias, three researchers were involved in the analysis, and we ensured that we achieved a reliable agreement between the researchers (i.e., with a Cohen’s kappa of 0.81). In addition, our manual labelling of the topics resulting from the topic modeling may be biased. For validating the topics generated from topic modeling, we consulted with a group of app developers, mainly focused on cross-platform (such as react native¹³ - an open-source software framework by Meta for Android and iOS mobile application development and flutter¹⁴ - an

¹³<https://reactnative.dev/>

¹⁴<https://flutter.dev/>

open source software development toolkit by Google for Android and iOS mobile application development). They were provided with representative words and 5 examples of each topic to validate the topic labels. They were asked to rate Y if they agreed with the label and N if they did not. The results show that they agree with our labels with 100% positive feedback.

4.6 Chapter Summary

This chapter studies the characteristics of a large number (2,058) of sports apps in the Google Play store and their user reviews. Through a manual analysis, we identified sports apps that cover 16 sports types (e.g., Football, Cricket, Baseball) and 15 main functionalities (e.g., Betting, Betting Tips, Training, Tracking). Through automated topic modeling, we extracted 14 topics from the user reviews, among which 3 are specific to sports apps (*accuracy of prediction*, *up-to-dateness*, and *precision of tools*). Finally, through sentiment analysis, we observed that users are mainly complaining about the advertisements and quality (e.g., bugs, content quality, streaming quality) of sports apps. Through the large-scale study focusing on a single app category, we demonstrate that analyzing a targeted category of apps (e.g., sports apps) can provide more specific insights than analyzing apps across different categories while still being relevant for a large number of apps. Besides, as a rapid-growing and competitive market, sports apps provide rich opportunities for future research, for example, to study the integration of data science or machine learning techniques in software applications or to study the factors that influence the competitiveness of the apps. We encourage future work to perform more studies on focused app categories to provide more meaningful observations and recommendations for the specific app categories.

CHAPTER 5 HOW CAN WE EXPLAIN USER RATINGS OF MOBILE APPS? A MULTI-DIMENSION STUDY OF ANDROID SPORTS APPS.

The Android mobile app store is a highly competitive marketplace, with a staggering 2.67 million apps available to users. In order to improve their ratings and gain a competitive edge, developers must understand the game ground among their competitors. We propose a three-dimension system analysis (market, categorical, and functional) rather than a global-dimension analysis to help developers understand the factors that impact ratings. The market dimension refers to the general apps in Google Play Store, the categorical dimension refers to a specific category of apps available in the app store (e.g., sports), and the functional dimension refers to a group of apps within a category that has similar functionalities (e.g., football betting apps in the sports category). In this study, we evaluate over 12,000 apps using the three-dimension data stratification and examine how the features from Google Play can explain app ratings at different dimensions. Our analysis indicates that the functional dimensions of the app explain the models better than the higher dimensions (such as the category or genre of the app). This means that focusing on the functional aspects of the app is more important in predicting app ratings than considering broader factors such as the app category. Furthermore, our multidimension approach allows us to identify dominant features specific to each category, meaning that we can identify the features that are most important for a particular type of app. By analyzing the importance scores of these features at the functional dimensions, we can determine which features are most critical for improving app ratings and gaining a competitive advantage in the market. Additionally, functional dimension analysis helps to understand the important features to win over competitors in a tight competition landscape (when ratings are very close). We conclude that research on the multidimension approach can help developers better understand unique factors that impact app ratings, enabling them to identify specific metrics to improve ratings over their competitors.

5.1 Introduction

In recent years, the mobile application development market has experienced exponential growth, with projected revenues exceeding 500 billion dollars and a growth rate of 12.8% [92]. App stores have become more than mere distribution platforms, as they now serve as a primary source of user feedback and are essential in determining the success of mobile applications. In the highly competitive mobile application market, developers face significant

challenges in making their applications stand out among millions of others. User acquisition and retention are critical metrics that determine an app’s success, making it essential for developers to focus on app store optimization (ASO) strategies to improve their application’s discoverability and attract more users [93].

The Google Play Store is the most widely used app store globally, offering over 2.7 million apps for download. However, it has not disclosed its metrics for calculating application rankings. Several studies have investigated the impact of ratings and reviews on an application’s success by shaping users’ decision to download the app [94]. These studies have established that highly rated applications are more likely to be preferred by mobile users. Specifically, it has been found that approximately 85% of users tend to abandon applications with a rating of two stars or less. Consequently, a comprehensive comprehension of the factors that contribute to improving an app’s rating is vital for effective user acquisition strategies.

Although app ratings are a crucial factor in ASO [5], there needs to be more clarity on how to influence them. Developers have limited control over certain factors that affect app ratings, such as the number of installs and user feedback. Nevertheless, they can optimize the metadata they control, such as the app’s title, screenshots, and description, to improve their app’s discoverability in the app store [93].

In order to enhance application ratings and optimize metadata, developers may benefit from a multidimension analysis that examines unique features and requirements of specific app categories. Our study adopts a three-dimension approach that explores the market, categorical, and functional dimensions. Such an approach generates tailored insights by identifying distinct features and behaviours for specific app functions and patterns and trends particular to specific categories. For example, while permissions related to a wallpaper (“set wallpaper”) are only specific to personalized app categories, such permissions would not be identified for general apps. Insights gleaned from functional and categorical dimension models can enable developers to design customized strategies that cater to the specific requirements of their target audience. Consequently, a more engaging and effective app experience can improve user satisfaction and ratings. By analyzing the app domain at functional and categorical dimensions, developers can make more informed decisions regarding feature prioritization, resource allocation, and marketing strategies. This can result in more efficient and effective development processes and improved app performance. The importance of multidimension analysis in generating such advantages has been highlighted in previous research [26].

Our study prioritizes both a) on-metadata, which the publisher has full control of, such as the app icon, category, video previews, etc., b) and off-metadata, which are uncontrollable, such as price, installs, and app permissions [17, 95] This paper aims to provide valuable in-

sights into the impact of on-metadata and off-metadata factors on app ratings, focusing on a 3-dimension analysis. We will examine the unique requirements of an app in the mobile app store (market dimension) and the categories of sports apps and then personalized apps (categorical dimension) and how they can be leveraged to improve app discoverability, followed by analysis among competitors of sports apps (functional dimension) to attract more users. Through this research, we aim to provide developers with a better understanding of the factors that affect app ratings and how to optimize their application’s metadata to achieve higher rankings in app stores. The selection of app categories for this study is based on several factors. Firstly, the sports app category is chosen due to its consistently increasing fan base in sports and fitness as a lifestyle, which has been observed in both general trends and within the Google Play Store. Additionally, sports apps encompass a wide range of functionality, such as social networks, sports tools, applications requiring real-time data updates, and predictive and analytical models. Thus, a study on this genre can provide valuable insights into the nuances of functionality at the functional dimension. We have defined our motivation to the sports apps more detail in Chapter 4. Secondly, the personalization category is selected due to its high popularity among users, ranking among the top 10 categories in the entire Google Play Store with 107,379 apps and a remarkably high average rating of 4.33 out of all categories. Therefore, this category is chosen as a reference for category-to-market study analysis. For this study, we focus on sports apps for our multidimension analysis, where as the personalization category will be a reference for category-to-market study analysis. Our research aims to comprehensively examine the factors influencing app rate prediction at a different dimension with a varying granularity of data.

Following the acquisition of data, we carried out a meticulous cleaning process, as well as redundancy and correlation analysis, on a sample of 12709 general apps, 1457 sports apps, and 1586 Personalization apps. For personalization apps, we selected the most frequent apps general app category. For the sports apps, we employed the methodology suggested by Chembakottu et al. [96]. The app rating prediction task was accomplished using a 3-dimension system that comprises the overall Google Play Store apps (Market dimension), a specific sports category (category dimension), and sports app competitors, such as football betting apps (Functional dimension), as elaborated in Section 5.3.1. At each dimension, we compared the performance of various regression models. There have been studies to understand the relationship between different features [9, 10, 50, 51]. We are considering the potential features as an explanatory variable predicting the ratings. The explanatory power as the impact factor of the factor. We analyzed the features affecting the app rating for different dimensions of granularity mentioned above. At the functional dimension, we analyze the features for relative ranking among competitors.. Specifically, our work aims to

answer the following three research questions (RQs)

• **RQ1: How do different dimensions of models perform in explaining the app ratings?**

The impact of data granularity and complexity on the model’s explainability is crucial in understanding how multidimension model performance can be leveraged to predict app ratings. In this study, we constructed regression models based on 29 extracted features as explanatory variables to the dependent variable rating, employing various techniques to identify the optimal model for rating prediction. Specifically, we employed the Case-Based Reasoning (CBR) model utilizing Euclidean distance to forecast new cases based on prior ones, as well as Random Forest Regression (RFR), Bagging Regressor (BR), and eXtreme Gradient Boosting (XGB) models. A Dummy Regressor (DR) served as a benchmark to compare model performance. Using normalized error metrics and a common test set to compare the results of multiple models, we evaluated the performance of each model within and across dimensions. The findings showed that the Random Forest model consistently performed well across models, and that model performance was better for lower dimensions than higher dimensions.

• **RQ2: What are the most influential features for predicting ratings for each dimension?** To improve app ratings, it is essential to understand the factors that impact ratings. In this study, we examined the coefficient values of the regression model defined earlier to identify the most influential factors. After determining the optimal model for rating prediction, we investigated the impact of individual features on prediction accuracy. To gain a more comprehensive understanding of feature importance, we conducted a three-dimension granularity analysis to assess whether the importance of features differed across the general market and categorical dimensions, such as sports and personalization. Our analysis identified that the feature importance at the functional dimension differs from those at the market dimension. Consequently, to achieve higher ratings among applications within the same dimension, we need to improve the factors specific to that dimension. Our findings suggest that a multidimension approach can be useful in identifying unique features that contribute to app ratings and in developing targeted strategies to improve ratings within functional dimensions and categorical dimensions.

• **RQ3: What factor influences the ranking of apps among competitors?** When faced with a market saturated with many applications offering similar functionality and ratings very close to each other, absolute rating prediction and feature identification become challenging. We propose that the relative ranking of each application within its specific functionality is calculated and analyzed based on its absolute rating. Relative ranking facilitates

a comprehensive evaluation of the benchmark of the application’s performance against its competitors, allowing for an objective assessment of its influencing features. It handles the skewness of the distribution and improves the interpretability of the models among competitors. Our results proved that when the competition is high among applications, the relative ranking helps to identify the crucial factors very specific to the functional dimension.

We analyzed each research question to address the null hypotheses as follows: For RQ1, we tested the hypothesis as specified in H_0 : *There is no significant difference in the performance of different regression models in predicting ratings across the three dimensions*. We utilized the K fold regression method for the ensemble models with $k=10$, while for the CBR model, we analyzed the model with K values (neighbours) from 1-20, choosing the model with the lowest error value. For RQ2, we tested the hypothesis as specified in H_0 : *There is no significant difference in the influence of different features on predicting ratings for each dimension*. To analyze this null hypothesis, we selected the best-performing model for each dimension and employed the Scott-Knott feature ranking method [97], a non-parametric approach that ranks features based on their effect size on rating values, to refine our understanding of feature importance. Finally, for RQ3, we tested the hypothesis as specified in H_0 : *There is no significant relationship between the features of an app and its ranking among competitors*. We followed the feature importance analysis for the previous null hypothesis to investigate this research question.

Chapter organization. The remainder of the chapter is organized as follows. Section 5.2 describes the experimental setup of our study. Section 5.3 describes our results from the experiments conducted for each research question. Section 5.4 discusses the implications of our findings. The threats to the validity of our findings are discussed in section 5.5. Finally, we conclude our chapter in section 5.6.

5.2 Experiment Setup

Figure 5.1 shows an overview of our study. This section explains our process of identifying, collecting, and preparing the data for answering our research questions. The detailed approaches used for answering each RQ are described in Section 5.3.

5.2.1 A: Data Gathering

In this study, the initial step involved the creation of a dataset. To achieve this, the Google Play Store was selected based on its popularity and easiness in data extraction, as indicated in previous works [98, 99]. Our study involves data crawling from the highly competitive

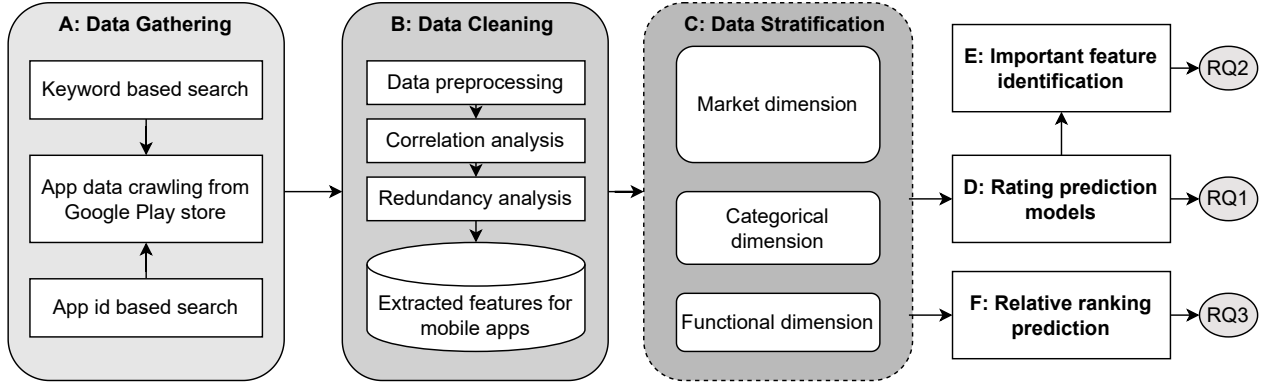


Figure 5.1 Overview of our study

Google Play Store, as evidenced by McIlroy and Hussain [98].

A prior investigation by Chembakottu et al. [96] demonstrated that the app results acquired from a keyword search on the Google Play Store are distinct from the outcomes produced by the genre filter. To overcome this limitation, the authors proposed an iterative keyword search approach that generates a set of keywords to search for in the description of the resulting apps to obtain apps that correspond to a specific category. In our study, we use an ensembled approach of the keyword-based search for identifying the detailed data collection of applications in a specific category. App ID-based search, which extracts databased on the IDs provided, will help to update the information in an existing dataset.

Keyword-based search: In our study, we utilized same set of keywords mentioned in paper by Chembakottu et al. [96] to extract the relevant apps from the Google Play Store. However, we employed the crawling tool previously mentioned to ensure that the data remained consistent and was up-to-date.

We extracted our data from the google play store on 4th March, 2023. We have extracted 12709 general apps and 1457 sports apps with the 29 features (Table 5.1) that we think are determining factors of application rating. Google play store has limitations to distributing the apps all over the world. Due to privacy and policies and many other reasons, the supported locations for distribution to google play users change [100]. We limited our country of choice as *US* and language as *English* for crawling our data. The methods of our data extraction process will be available in our replication package; however, due to the elicitation process, the data might be different from what we are using. For reproducibility, we will support the data along with this work. We extracted the permission details associated with applications along with metadata that we can see on the google play website¹.

¹<https://play.google.com/store/apps>

App ID based search: In the preliminary phase of dataset generation, we utilized application identification numbers (app IDs) provided by Mohsen et al. [101]. This dataset was selected due to its comprehensive coverage of 350,000 mobile apps from the Google Play Store in 2020, including features and characteristics of apps that were removed from the store. The use of this dataset provides a foundation for replicating previous studies while also updating the dataset with new information. As the previous study is relatively recent compared to our work, updating the data included more information in the dataset. We used the app IDs from the paper and passed them to the Google-play-scraper Python library² to scrape the latest and most comprehensive app descriptions and information. This methodology was implemented to ensure that the apps analyzed were current and not obsolete due to the Google Play Store’s data removal procedure, thus maintaining consistency within the dataset.

Feature name	Type	Definition	Rationale
Real Installs	N	The total number of installs that the app has	Indicates the success and popularity of the App.
Score	N	User rating score of the app in the range (0,5].	Indicates the app’s quality and may be used to compare it to other apps.
Ratings	N	The total number of user ratings the app has received.	Indicates the level of user engagement with the app.
Reviews	N	The total number of user reviews the app has received.	Provides detailed feedback on the user experience and can be used to identify areas for improvement.
Price	N	The price of the app	Can be used to compare pricing strategies with other apps and analyze the impact of pricing on user sentiment.

Continued on next page

²<https://pypi.org/project/google-play-scraper/>

Table 5.1 The extracted app features from Google Play Store. The data types are defined as N: Numerical, B: Boolean, C: Categorical.

continued from previous page

Feature name	Type	Definition	Rationale
Contains Ads	N	Indicates whether the app contains adss.	Can be used to analyze the impact of ads on user experience and sentiment.
Released	N	The year the app was released.	Provides insight into the app's lifespan and can be used to analyze trends in the mobile app market.
app Icon	B	The availability of icon	The app icon is the identity of the apps, and it can be pivotal for app positioning and downloads
Header Image	B	The availability of header image	The header image is an important visual element in the app store listing. It can convey the app's overview and provide more information about its features and functionality.
Screenshots	B	The number of screenshots available in the app store listing	It demonstrates the app's features and functionality. It acts as a static demo of the apps, leading to more downloads and higher ratings.
Video	B	The availability of a video for the app	It is a demo of the app's features and functionalities, which can lead to a better understanding and more downloads.
Video Thumbnail	B	The availability of the video thumbnail image file	The video thumbnail attract users to watch the video.
Content Rating	C	The content rating of the app	The content rating indicates the level of maturity required to use the app; it contributes to the app's visibility and downloads.

Continued on next page

continued from previous page

Feature name	Type	Definition	Rationale
Ad Supported	B	The availability of ads in the app	The presence of ads in the app can contribute the users' experience and perception of the app, which can lead to lower ratings and fewer downloads.
Last Updated Date	N	The year when the app was last updated	The last updated date year can provide important information about the app's maintenance practices, contributing to the app's performance and ratings.
Camera Permissions	C	The app's access to the device camera	The camera permission is an important security and privacy feature, and it can contribute to the users' trust and perception of the app.
Phone Permissions	C	The app's access to the device's telephony features	The phone permission is an important security and privacy feature, and it can contribute to the users' trust and perception of the app.
Device ID and Call Information	C	It provides information about the mobile device ID and call information	It is useful for understanding how an app uses the unique device ID and calls information, which can help assess privacy and security risks associated with the app.
Storage	C	It provides information about the mobile device's storage capacity and usage	It is important for understanding how much storage an app uses, which can help users manage their device's storage capacity and assess the app's resource usage.

Continued on next page

continued from previous page

Feature name	Type	Definition	Rationale
Wi-Fi Connection Information	C	It provides information about the mobile device's Wi-Fi connection status and configuration	It is important for understanding how an app uses Wi-Fi connectivity, which can help assess the app's functionality and resource usage.
Photos, Media and Files	C	It provides information about the mobile device's photos, media, and files	It is useful for understanding how an app uses the user's media and files, which can help assess privacy and security risks associated with the app.
Microphone	C	It provides information about the mobile device's microphone	It is important for understanding how an app uses the user's microphone, which can help assess privacy and security risks associated with the app.
Other	C	It provides information about other features or permissions used by the app	It is useful for understanding any other features or permissions used by the app that are not included in the other categories.
Uncategorized	C	It provides information about any features or permissions used by the app	It is useful for understanding any other features or permissions used by the app that are not included in the other categories.
Length of Description	N	It provides the length of the app's description	It can help in understanding how app descriptions are written and can help identify correlations between app descriptions and user ratings.
Length of Name	N	It provides the length of the app's name	It can help in understanding how app names are written and can help identify correlations between app names and user ratings.

Continued on next page

continued from previous page

Feature name	Type	Definition	Rationale
Identity Per- missions	C	The app's access to the device's identity information	The identity permission is an important security and privacy feature, and it can contribute the users' trust and perception of the app.
Contacts Per- missions	C	The app's access to the device's contact information	The contacts permission is an important security and privacy feature, and it can contribute the users' trust and perception of the app.
Offers in-app purchases	N	Indicates if the app offers in-app purchases.	Provides insights into the app's monetization strategy and revenue model.
Genre	C	The category or genre of the app.	Provides an understanding of the app's target market and user preferences.

5.2.2 B: Data Cleaning

Data preprocessing: The quality of data is crucial for optimal model performance. In this study, data was extracted from Google Playstore, consisting of both numerical and textual values. To begin the preprocessing phase, applications with zero rating scores were removed. Basic text cleaning techniques were then applied using a regular expression library in Python³. Specifically, we aimed to eliminate noise from the app descriptions and names, including typing errors, special characters, emojis, and stopwords, using the Natural Language Toolkit (NLTK) prior to computing the length of text. The length serves as a quantitative metric for our analysis. Additionally, we manually manipulated certain features such as the release and update dates, retaining only the year. On-metadata pertaining to media elements, such as screenshots and header images, were evaluated by counting the number of screenshots and by assessing their availability indicated by binary variables (0s and 1s). For metadata relating to app permissions, we considered all available permission types at the time of data collection, and treated them as categorical variables in our analyses.

³<https://docs.python.org/3/library/re.html>

Correlation analysis: Prior to constructing our predictive models, we conducted a correlation analysis on the features to identify potential confounding variables that could affect the relationship between the predictor and response variables. To accomplish this task, we employed the Spearman correlation test, which is preferred due to its ability to handle non-normally distributed data [102, 103]. To identify highly correlated variables that may cause multicollinearity issues in our models, we set a correlation value threshold of $\rho > 0.7$. By doing so, we can make informed decisions on how to handle these variables to ensure the accuracy and robustness of our models.

Redundancy analysis: We conducted a redundancy analysis to identify potential variables that could impede the relationship between the explanatory variables and app ratings. To detect redundant variables, we fitted preliminary models to a linear regression model that explains each explanatory variable using the other explanatory variables. We assessed the explanatory power of each variable using the R^2 statistic, which measures the proportion of variance explained by the model. We used a linear regression model and set a threshold of 0.9 to exclude redundant variables based on prior research by McIntosh et al. [103]

5.2.3 C: Data Stratification

A central objective of our research is to investigate the factors that influence application ratings across multiple dimensions of analysis. To achieve this goal, we conducted a data stratification process based on three dimensions: Market, Category, and Functional (Figure 5.1). At the Market dimension, we aggregated data on general apps from the Google Play Store that we obtained through web crawling. At the Category dimension, we focused on two categories: (i) the most common genre from the Market dimension data, namely *Personalization*, which includes apps that enable users to customize and personalize their Android devices by changing themes, wallpapers, icons, fonts, and other visual elements; and (ii) the sports category. However, given that developers classify their apps into genres based on their intuitions (i.e., on-metadata), the likelihood that apps that need to be under a given category might not be tagged in the same genre is high. To ensure a more inclusive dataset of sports apps, we adopted the keyword-based search approach proposed in [96], which involved a two-stage process of automatic filtering followed by a manual cleaning process to eliminate any games within the sports category. For the functional dimension data analysis in this study, we followed the label names mentioned in previous work [96]. Our approach involved a combination of automated and manual analysis. Firstly, we grouped the apps that had the label names mentioned in their descriptions into their respective categories using automated methods. Next, we performed a manual analysis on the dataset to verify and modify

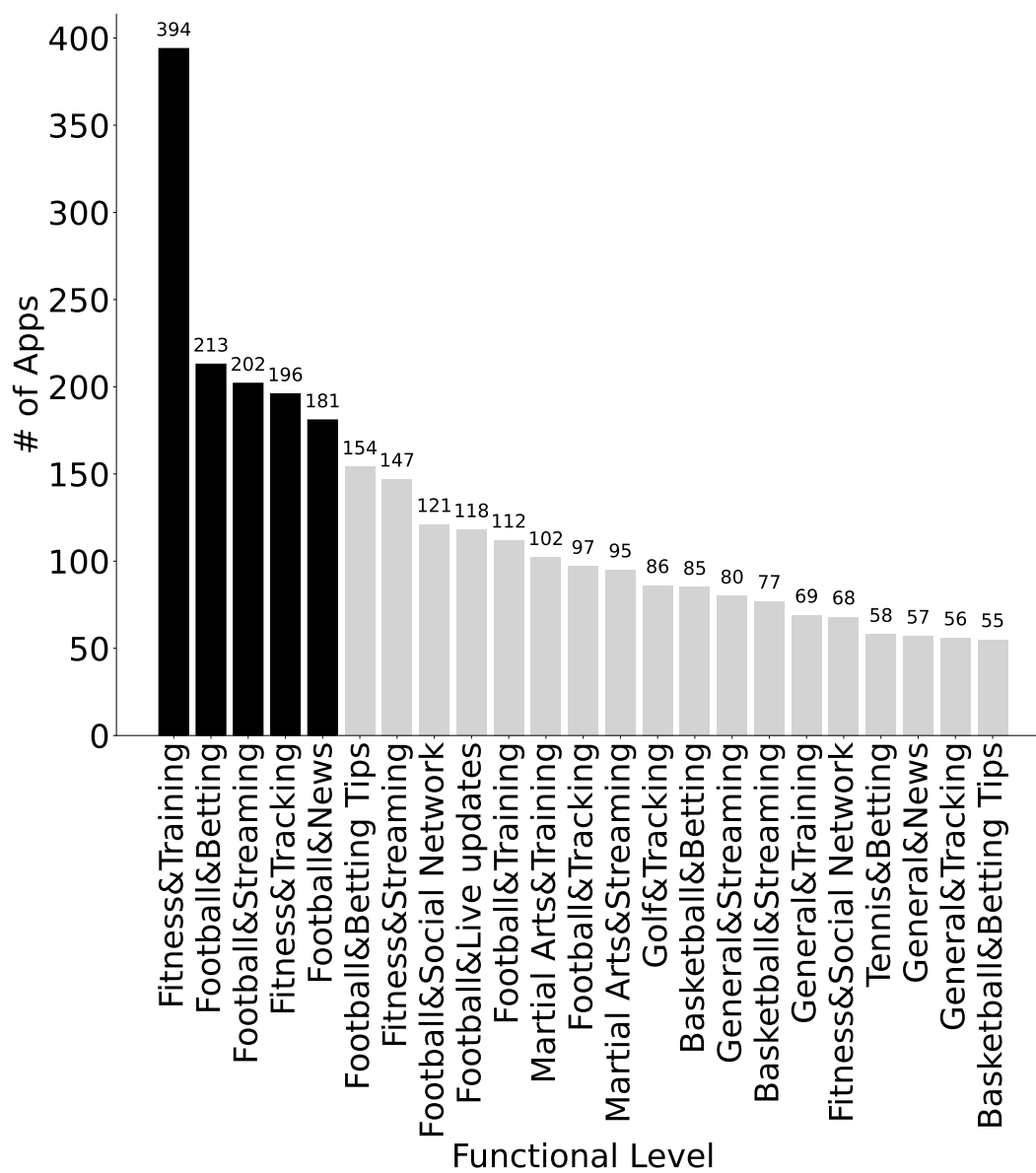


Figure 5.2 Distribution of different functionalities in sports category

the results. Figure 5.2 provides the distribution of the functional dimension apps obtained through this process. To streamline our analysis, we have selected and focused on the top 5 functionalities for the remainder of the study. This approach allows us to delve deeper into the nuances of these functionalities and derive meaningful insights.

5.2.4 E: Rating prediction models

Regression models: Upon finalizing our dataset, our goal is to apply different regression models to the data. The selection of the regression models is based on their repeated use in previous regression analysis studies and their interpretability [104]. To this end, we employ the Case-Based Reasoning (CBR) model, a well-established problem-solving approach that has found extensive usage in software engineering research. Previous studies have reported on the successful implementation of CBR in fields such as customer relations, mining, and forecasting[9, 50, 105]. The CBR model involves reasoning from past experiences to solve new problems. Additionally, we utilize three ensemble decision tree methods: Bagging Regressor (BR)[11], Random Forest Regressor (RFG)[10, 11, 106], and XGBoost Regressor (XGB)[11]. These models recursively split the features based on the lowest variance to predict ratings. The performance of these models is compared to a Dummy Regressor (DR) model[9], where ratings are randomly selected from a predetermined set of values. We assume the CBR and ensemble decision tree models will outperform the DR model.

Model settings: In our study, we utilized the methods proposed by Sarro et al. [9] for the CBR model. Before selecting the model for evaluation, we analyzed K-neighbour values ranging from 1 to 20. We computed the mean and median absolute residual for each model as outlined in the literature. Furthermore, we employed the Wilcoxon signed rank test[107] to obtain statistics and p-values for each error measure, enabling us to determine the model's statistical significance for various K values. For all other models, we utilized the scikit-learn library⁴ for cross-validation and evaluation methods. We have also documented all hyperparameters utilized for models and cross-validation in our replication package. We have provided the necessary details for replicating our code.

Evaluation Criteria and Validation Methods: Defining a set of evaluation criteria is necessary in our case as we compare multiple modes. To evaluate which regression model works better we use six common metrics:

⁴<https://scikit-learn.org/>

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5.1)$$

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5.2)$$

Root Mean Squared Error (RMSE) [108, 109]

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5.3)$$

Normalized Root Mean Squared Error (NRMSE)

$$NRMSE-SD = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\sigma}, NRMSE-IQR = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{Q1 - Q3} \quad (5.4)$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5.5)$$

In the above equations, n is the number of observations, y_i is the actual value of the target variable for the i -th observation, $\bar{y}$ is the mean of the observed response values, and $\hat{y}_i$ is the predicted value of the target variable for the i -th observation. To evaluate the difference between the predicted and actual values of the target variable, we utilized three metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error ($RMSE$). The MSE (Equation 5.1) measures the average squared difference, the MAE (Equation 5.2) measures the average absolute difference, and the $RMSE$ (Equation 5.3) measures the square root of the average squared difference. $NRMSE$ (Equation 5.4) [110] it is an extension of the RMSE that normalizes the error by the standard deviation (σ) and interquartile range (Q1-Q3) of the data. NRMSE provides a normalized measure of error that allows for easier comparison between different datasets or models, as it considers the range of values in the target variable. This is particularly useful in our case as we compare models predicting rating with the different test sets. The lower the value of each metric, the better the performance of the regression model. The R^2 (Equation 5.5) score measures the

proportion of variance in the response variable that is explained by the predictor variables. A higher R^2 score indicates a better fit of the model to the data.

5.2.5 D: Important feature identification

Feature importance score: Once we identified the best model for each dimension of our analysis, we selected the best model to identify the features that affected the rating prediction for each dimension across three dimensions. For example, for Random Forest Regressor model, to evaluate the importance of each feature for rating prediction. The feature importance was calculated by evaluating the impurity index of the features and then normalizing the scores. The features were sorted based on their importance scores, which were used to rank the features by their contribution to the model's performance [111].

Feature ranking: To identify the statistically significant differences in the feature importance scores, we used the Scott-Knott Effect Size Difference (ESD) algorithm [97]. This algorithm evaluates the variance of the importance of features in each possible split. It selects the split that reduces the variance within each group and maximizes the variance between groups. By grouping together similar features, the ESD algorithm helps identify the most important features for rating prediction and provides a ranking of features based on their relative importance.

5.2.6 F: Relative ranking prediction

The relative ranking is necessary for rating prediction because it considers the competitive landscape and considers the entire range of ratings for each functionality. Absolute rating models are vulnerable to outliers and rating bombardment phenomena[112], which can distort the ratings of apps and lead to inaccurate predictions. The relative ranking prediction has focused mainly on the competitors (at the functional dimension) to understand more trends and patterns and the important features affecting the rank among competitors. Relative ranking can be used to reduce the biases in the data and have a more objective performance measure. The relative ranking can be calculated in Equation 5.6 [113] Let's assume we our absolute rating is $X = \{x_1, x_2, \dots, x_n\}$ with n elements. We want to rank these elements from the smallest to the largest, and we will use the 'max' method for handling ties.

1. Sort the dataset in ascending order, so we get $X_{sorted} = \{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$, where $x_{(i)}$ denotes the i^{th} smallest element in the dataset.
2. Assign a rank r_i to each element x_i in the dataset as follows:

$$r_i = \begin{cases} i & \text{if } x_i \text{ is not tied} \\ \max(j) & \text{if } x_i \text{ is tied with } x_{(j)} \text{ for some } j \end{cases} \quad (5.6)$$

In this equation, $\max(j)$ denotes the maximum rank value that can be assigned to the tied group of elements.

5.3 Experiment Results

In this section, we demonstrate our motivation, approach and results for our research questions

5.3.1 RQ1:How do regression models perform in predicting rating across three dimensions?

Motivation

Previous research has extensively discussed the significance of user-perceived information, such as ratings and reviews, in influencing the success of mobile applications [9, 10, 11, 26, 50, 94]. However, prior studies have been restricted to either the market dimension, i.e., limited to the app stores in which the research was conducted or the genres provided by the app stores along with the market dimension. It has been observed that metrics may have nuances when analyzed more granularly [114]. Therefore, in this research question (RQ), we aim to analyze whether regression models are consistent across three dimensions of analysis and whether the performance varies across the dimensions or not.

Approach

In addition to analyzing the impact of user-perceived information on the success of mobile applications, our research question aims to investigate the consistency of rating prediction models across different groups. Specifically, we plan to analyze multiple regression models to understand which model performs better within each group. Additionally, we aim to explore whether data stratification affects the performance of the models. Our primary motivation for conducting this research is to gain a comprehensive understanding of the nuances of rating prediction models and provide insights that can aid in developing effective user acquisition strategies. By analyzing the consistency of regression models across different groups and exploring the impact of data stratification, we aim to contribute to the literature on the topic and provide practical implications for app developers and marketers.

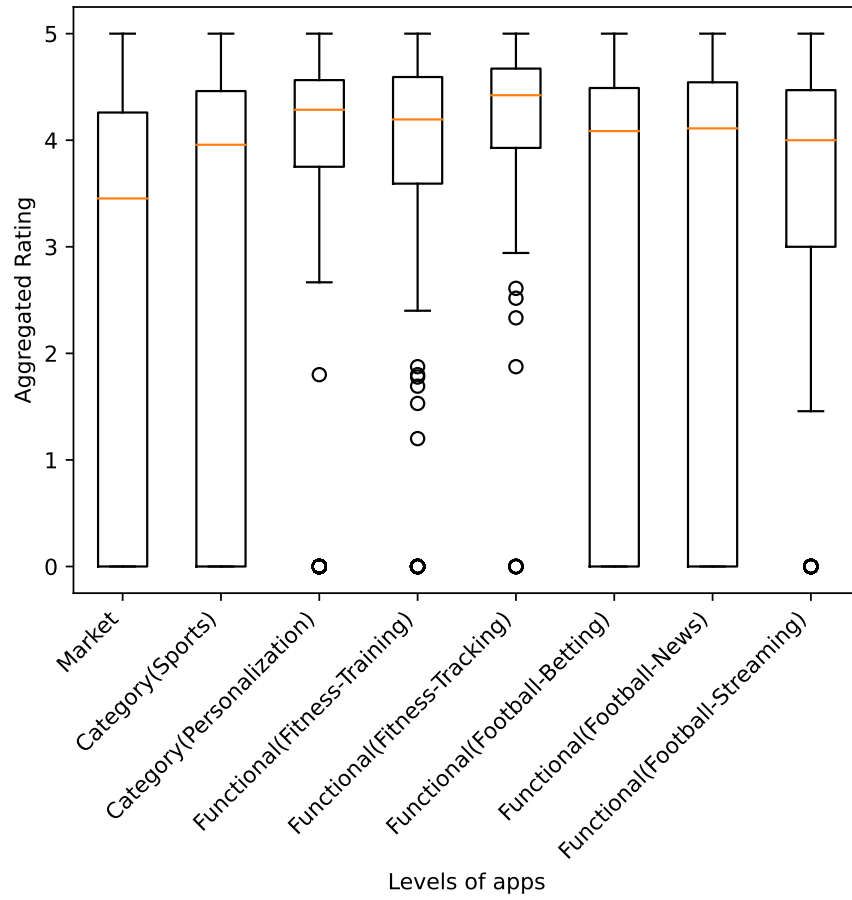


Figure 5.3 Distribution of average app rating for each data dimension. Each boxplot shows the distribution of the average rating of the all apps in the given dimension

Model evaluation within each dimension : We use the average rating provided by the Google Play Store as our dependent variable. This metric is calculated over the history of ratings provided by users since the apps were launched. Figure 5.3 displays the rating distribution at each dimension of analysis. Our main objective is to determine which regression model performs better for each dimension of data analyzed. We consider four models - Case-Based Reasoning (CBR) with the best k , Random Forest (RF), Bagging Regressor (BR), and XG Boost (XGB). The performance of each model is evaluated against a Dummy Regressor, which randomly selects outputs from the test dataset. We use three common evaluation metrics, namely Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), for our analysis. Since each dimension has a different test dataset, we use Normalized Root Mean Squared Error (NRMSE) with Standard Deviation (SD) and Interquartile Range (IQR) to normalize the data and make the results comparable [115]. Normalizing the data allows us to evaluate how well the rating score fits for different dimensions. The SD-based NRMSE represents the ratio between the variation not explained by the regression and the overall variation in rating [110]. On the other hand, IQR is less sensitive to extreme values. In order to choose the best k value for the CBR method[9], we evaluated MAE for values from 1-20 and choose the k values with the least MAE value. All of our models undergo K-fold cross-validation over the entire dataset where ($K=10$). The cross-validation is done in the way that the dataset is first divided into K equal folds. The model is then trained on each of the folds tested on the remaining fold. This process is repeated K times, with each fold used as the test set exactly once. Finally, the results of each fold combined and evaluated the metrics for the model. In this approach, the entire dataset is used for training and testing.

Model evaluation with common testsets: As illustrated in Fig. 5.3, the rating ranges differ for each dimension, leading to variations in error ranges for the models at each respective dimension. These discrepancies can be attributed to the inherent differences within each dimension's test set. In order to enable a fair and consistent evaluation of regression models across multiple dimensions, it is necessary to establish a standardized test dataset [108, 109]. Specifically, we have constructed a three-dimension dataset for sports apps and a two-dimension dataset for personalization apps. We have aggregated all functional groups for the sports apps and utilized the functional dimension data as our test data for the categorical (Sports) and general market dimensions. Care was taken to ensure that there were no instances of duplicated data at either the category or market dimension. Similarly, we lack a functional dimension for the personalized apps and therefore employ the category dimension as the standardized test set for the market dimension. By employing a common dataset for model performance analysis, the resulting evaluations are more reliable and enable a

rigorous assessment of the generalizability of the models across dimensions, as they are tested on unseen data [108]. We use three common evaluation metrics, namely, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) for our analysis. We don't have the Normalised metrics here because we use the common test set across the dimensions.

Results

Random Forest is the most effective model for regression tasks across different dimensions and categories, as it consistently outperformed other models The table 5.3 presents the performance of various regression models for each dimension considered, and the range of errors varies across categories due to differences in the distribution of the ratings, as illustrated in Figure 5.3. The results demonstrate that the Random Forest Regressor model consistently outperforms other models in most categories based on the evaluation metrics, such as MSE, MAE, and RMSE, which suggest that lower values correspond to better model performance. Based on the same metrics, the Bagging Regressor model also performs well in most categories. However, the Case-Based Reasoning model performs poorly across most categories, indicating that it may not be suitable for predicting the target variable. Additionally, the XG Boost Regressor model has higher NRMSE values in most cases, indicating lower accuracy in predicting the target variable relative to the variability of the data. It is worth noting that all selected regression models perform better than the baseline model, which was expected. To facilitate model comparison across different datasets, the data was normalized, and NRMSE was used to measure prediction accuracy relative to the variability in the data. Overall, the results demonstrate the importance of selecting appropriate regression models for different applications and considering the dimension and category of the apps to improve model performance. **Multidimension analysis of apps showed that the predictive models' performance decreased as the granularity dimension increased, with the lowest RMSE being observed at the functional dimension, followed by the category dimension and the market dimension.** In this study, performance analysis results are presented in Table 5.4 based on a hierarchical approach, in which lower-dimension data is used as a test set for higher dimensions. The results of the 3-dimension performance analysis for sports apps indicate that predictive models are better suited for functional-dimension predictions, as evidenced by the best-performing model, Random Forest, which achieved an RMSE of 0.54. In contrast, the error increased to 0.58 and 0.61 for the category and market dimensions, respectively. This trend is not limited to the best-performing model but is also visible in all regression models.

Table 5.3 Model performance across different levels while predicting absolute rating of apps

Level		Model	MSE	MAE	RMSE	NRMSE-SD	NRMSE-IQR
Market	General	Case Based Reasoning	0.26	0.38	0.51	0.94	0.76
		Bagging	0.23	0.35	0.48	0.88	0.71
		Random Forest	0.22	0.33	0.47	0.86	0.7
		XG Boost	0.42	0.57	0.69	0.99	0.89
		Dummy	0.82	0.89	0.98	1.8	1.2
Category	Sports	Case Based Reasoning	0.39	0.44	0.62	0.96	0.81
		Bagging	0.4	0.45	0.63	0.97	0.82
		Random Forest	0.36	0.43	0.6	0.92	0.78
		XG Boost	0.46	0.49	0.68	1.04	0.88
		Dummy	0.84	0.67	0.91	1.39	1.18
Category	Personalization	Case Based Reasoning	0.11	0.25	0.33	0.9	0.7
		Bagging	0.12	0.26	0.35	0.93	0.72
		Random Forest	0.11	0.25	0.33	0.9	0.7
		XG Boost	0.11	0.25	0.34	0.91	0.71
		Dummy	0.23	0.36	0.48	1.29	1
Functional	Fitness-Training	Case Based Reasoning	0.32	0.38	0.56	0.9	0.69
		Bagging	0.33	0.41	0.58	0.93	0.71
		Random Forest	0.26	0.36	0.51	0.82	0.63
		XG Boost	0.31	0.43	0.56	0.9	0.69
		Dummy	0.79	0.61	0.79	1.26	0.97
Functional	Football-Betting	Case Based Reasoning	0.26	0.37	0.51	1.01	0.98
		Bagging	0.3	0.39	0.55	1.09	1.06
		Random Forest	0.26	0.37	0.51	1.01	0.98
		XG Boost	0.41	0.52	0.64	1.28	1.23
		Dummy	0.44	0.5	0.66	1.32	1.27
Functional	Football-Streaming	Case Based Reasoning	0.43	0.49	0.66	1.06	0.87
		Bagging	0.4	0.49	0.63	1.02	0.83
		Random Forest	0.37	0.44	0.61	0.98	0.8
		XG Boost	0.56	0.59	0.78	1.21	1
		Dummy	0.96	0.72	0.94	1.86	1.75
Functional	Fitness-Tracking	Case Based Reasoning	0.28	0.36	0.53	1	0.85
		Bagging	0.28	0.37	0.53	0	0.84
		Random Forest	0.24	0.34	0.49	0.93	0.78
		XG Boost	0.48	0.62	0.69	0.99	0.89
		Dummy	0.87	0.95	0.98	1.8	1.2
Functional	Football-News	Case Based Reasoning	0.3	0.39	0.55	1.01	0.93
		Bagging	0.32	0.42	0.56	1.04	0.95
		Random Forest	0.3	0.39	0.55	1.01	0.93
		XG Boost	0.52	0.72	0.69	0.99	0.89
		Dummy	0.93	1	1.3	2.1	1.8

Table 5.4 Model performance across different dimensions when common testsets are used

Dimension		Model	MSE	MAE	RMSE
Market dimension	General	Case Based Reasoning	0.45	0.54	0.67
		Bagging	0.46	0.56	0.69
		Random Forest	0.41	0.5	0.61
		XG Boost	0.51	0.63	0.69
		Dummy	0.66	0.76	0.88
Category dimension	Sports	Case Based Reasoning	0.39	0.47	0.62
		Bagging	0.39	0.47	0.62
		Random Forest	0.34	0.44	0.58
		XG Boost	0.45	0.51	0.67
		Dummy	0.66	0.76	0.88
Functional dimension	Top 5 Functionalities	Case Based Reasoning	0.34	0.41	0.58
		Bagging	0.34	0.42	0.58
		Random Forest	0.29	0.38	0.54
		XG Boost	0.56	0.6	0.74
		Dummy	0.66	0.76	0.88

Dimension		Model	MSE	MAE	RMSE
Market dimension	General	Case Based Reasoning	0.17	0.32	0.41
		Bagging	0.21	0.36	0.46
		Random Forest	0.16	0.33	0.41
		XG Boost	0.2	0.35	0.45
		Dummy	0.23	0.36	0.48
Category dimension	Personalization	Case Based Reasoning	0.12	0.26	0.34
		Bagging	0.11	0.25	0.33
		Random Forest	0.1	0.24	0.33
		XG Boost	0.11	0.25	0.34
		Dummy	0.23	0.36	0.48

We performed a 2-dimension analysis for the personalization category, where the category dimension has higher performance than the market dimension. As a result, the null hypothesis for the research question, H_0 : *There is no significant difference in the performance of different regression models in predicting ratings across the three dimensions.*, is rejected.

The results of our analysis indicate that the Random Forest model regression exhibits consistent performance across various dimensions. Furthermore, we have deduced that using lower dimensions of data, trends are more granular, leading to higher model performance. Conversely, at higher dimensions, the granularity of the trends becomes more complex, resulting in reduced model performance.

5.3.2 RQ2: What are the most influential features for predicting ratings for each dimension?

Motivation

There exist varying dimensions of metadata, as illustrated in Table 5.1, for applications in the Google Play store. Determining the influential features that affect application ratings is crucial for developers to enhance the success of their apps. In Section 5.3.1, we could accurately predict application ratings while accounting for data variance across different dimensions. In this RQ, we aim to identify the features that contribute to obtaining a favourable score for an application. This feature identification will aid in revealing individual or groups of features that impact app ratings. Such granular information can facilitate rating improvement for applications within a specific search (functional dimension), a particular genre (category dimension), or the Google Play store itself (market dimension).

Approach

In our previous research question (RQ1), we observed that the Random Forest regression model consistently performed well across all dimensions of data analysis. Therefore, we have chosen this model as the focus of our current research question (RQ2). Random Forests are a type of ensemble model that is known to be explainable[104]. Using the Random Forest model, we can easily identify errors, correlations, and feature importance[116]. For our analysis, we are using the sci-kit-learn implementation⁵ of the Random Forest algorithm. This implementation enables us to identify the features and their importance scores based on the impurity index. In the previous section (RQ1), we defined a parameter grid for the Random Forest algorithm and conducted a grid search to tune the hyperparameters. The main focus of our parameter grid is on the depth of the trees (*max_depth*), the number of estimators (*n_estimators*), and the number of features considered for the split (*max_features*). Once we obtain the features and their importance scores based on the model, we apply the Scott-Knott ESD (Effect Size Difference) algorithm[97] to rank the features. The Scott-Knott ESD algorithm ranks the features based on their importance scores and groups them into subsets with statistically different importance scores. This helps us to identify the most significant features of the given model. For analyzing this research question, we focus on the top 5 groups from Scott-Knott ESD as the ranking of the features for each dimension.

⁵<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

Results

Features of Screenshots, Reviews, and Length of Description consistently emerge as important across categories. Table 5.5 displays the top five ranked features for each dimension, as determined by Scott-Knott Effect Size Difference (ESD) analysis. The Scott-Knott ESD methodology groups together feature that demonstrates equivalent predictive significance. Upon analyzing each category’s top five feature groups, we observe that certain features such as Screenshots, Reviews, and Length of Description repeatedly appear across dimensions. The presence of features across multiple dimensions indicates that these features are influential and significant in determining the outcome of the analysis, regardless of the dimension of granularity [114]. It suggests that these features have a consistent impact on the rating score across the market, categorical and functional dimensions of the data hierarchy. This consistency indicates the robustness of these features and their importance in the overall model. Identifying such features helps app developers to put their focus on improving them in the Google Play Store.

The observation that Screenshots consistently emerge as significant across categories underscores their importance in enhancing app ratings at all dimensions. This identification highlights the value of adhering to the Google Console guidelines [117] for adding screenshots to improve app ratings. By paying close attention to these guidelines, developers can optimize their app’s visual presentation, increasing its appeal and engagement for potential users and ultimately improving its overall rating. We can draw a similar conclusion for the Length of the Description, as it also falls under on-metadata, the data category accessible to developers on the Google Play Store.

In contrast, Reviews constitute a distinct category of user-perceived information we analyzed in Chapter 4. During the Market dimension analysis, we observed that the feature *Genre* ranked first in terms of significance, thereby highlighting the importance of conducting a granular dimension analysis to obtain deeper insights and derive meaningful implications for the optimization of application metadata and improvement of app ratings.

Comparison of two categories revealed unique features specific to category dimension not present in the market dimension. The results presented in Table 5.5 demonstrate a comparison of two genres in the category dimension, namely sports and personalized apps. For sports apps, we identified that features such as *view wifi connections* and *precise location (GPS and network-based)* were ranked in the top three. This is due to the fact that sports apps often require real-time data updates and interactive features that rely on network and location services.

Table 5.5 Feature ranking using Scott Knott across different dimensions of apps

Dimension		Features	Rank
Market	General	Genre	1
		Reviews	2
		Length of Name, Length of Description	3
		Updated	4
		View wifi connections, Run at startup, Screenshots	5
Category	Sports	Screenshots	1
		Updated, Reviews, Length of Description	2
		View wifi connections, Length of Name,	3
		precise location (GPS and network-based)	4
		,run at startup, offersIAP, control vibration videoImage, pair with Bluetooth devices ContainsAds, ContentRating, Genre	5
Category	Personalization	Length of Description	1
		Length of Name, Reviews	2
		run at startup, receive data from Internet, screenshots	3
		modify system settings, set wallpaper, videoImage, Google Play license check,	4
		prevent device from sleeping, contentRating containsAds, view network connections	5
Functional	Fitness-Training	Screenshots	1
		Length of Description	2
		prevent device from sleeping, reviews	3
		control vibration, Length of Name	4
		run at startup, containsAds, view Wi-Fi connections, read calendar events plus confidential information, updated	5
Functional	Football-Betting	Length of Description	1
		Reviews	2
		Updated, Ratings	3
		contentRating, offersIAP, screenshots	4
		change your audio settings,change/intercept network settings and traffic, connect and disconnect from Wi-Fi, modify system settings, Length of Name	5
Functional	Football-Streaming	Screenshots	1
		Reviews	2
		ContentRating	3
		prevent device from sleeping, Length of Description	4
		Control vibration, Length of Name	5
Functional	Fitness-Tracking	Reviews	1
		Screenshots	2
		contentRating, Length of Description	3
		Updated	4
		offersIAP, pair with Bluetooth devices, Length of Name,	5
Functional	Football-News	Screenshots	1
		Reviews	2
		ContentRating, Length of Description	3
		Updated	4
		control vibrations	5

Note: Each rank contains features that are statistically indistinguishable from each other in terms of their importance score.

Additionally, we found that the feature *Genre* is also an important factor for sports apps, underscoring the need to analyze the functional dimension in such studies.

Comparison of two categories revealed unique features specific to category dimension not present in the market dimension. The results presented in Table 5.5 demonstrate a comparison of two genres in the category dimension, namely sports and personalized apps. For sports apps, we identified that features such as *view wifi connections* and *precise location (GPS and network-based)* were ranked in the top three. This is due to the fact that sports apps often require real-time data updates and interactive features that rely on network and location services. Additionally, we found that the feature *Genre* is also an important factor for sports apps, underscoring the need to analyze the functional dimension in such studies.

In contrast, the most important features for personalized apps were permissions related to *receive data from the Internet, set wallpaper, and modify settings*. These features were unique to the category and did not appear in any other dimension of analysis. Overall, our findings suggest that analyzing different dimensions of detail is crucial for deriving meaningful insights and developing tailored strategies that cater to the specific requirements of each app category.

The prioritization of network and traffic in football betting apps is relatively high, while fitness tracking applications predominantly emphasize the functional dimension details of Bluetooth devices. Multidimension analysis has facilitated the identification of dimension-specific features that necessitate consideration for obtaining favourable ratings and informing data-driven business decisions. Comparing category-dimension apps to functional-dimension apps, specific features are identified for each functionality. For example, football betting apps are time-sensitive. They require real-time updates, making network connections a high-priority feature such as *change/intercept network settings and traffic, connect and disconnect from Wi-Fi*. Fitness tracking apps focus on connecting and tracking activities with wearable devices, making the important feature *pair with Bluetooth devices*. Football streaming apps focus on providing an uninterrupted visual experience to the user, making the feature *prevent the device from sleeping* important. The highlighted important features at the functional dimension are not identified in the category or general dimensions. As a result, the null hypothesis for the research question, H_0 : *There is no significant difference in the influence of different features on predicting ratings for each dimension.*, is rejected.

Multidimension analysis reveals crucial features impacting app ratings, with Screenshots, Reviews, and Length of Description being consistently important. Comparing sports and personalized apps highlights unique features at the category dimension. Functional dimensions analysis put forward the need for tailored strategies and data-driven decisions for app optimization.

5.3.3 RQ3: What factor influence the ranking of apps among competitors?

Motivation

Our study demonstrated that analyzing functional-dimension applications unveils additional function-specific information. However, we observed varying competitive landscapes for different functionalities (as shown in Figure 5.3). Fitness-tracking displays a highly competitive environment, while Football-Betting and Football-News exhibit more substantial variations in app quality and user experiences. These differences in competitive ranges imply that distinct prioritization strategies and user expectations are required for each functionality. Employing relative ranking for rating prediction and feature importance analysis enables the derivation of more contextualized insights among competitors. This approach is less vulnerable to outliers than absolute values and more resilient to rating bombardment phenomena [112]. Relative ranking models provide more information than absolute rating models in competitive landscapes because they consider the entire range of ratings for each functionality rather than just the raw scores. Using a relative ranking approach, it is possible to identify which apps are performing better or worse than their competitors within a specific functionality and which features drive this performance difference.

Approach

In this research question, our focus is solely on analyzing the competitors of sports apps at the functional dimension. Consequently, we have limited our dataset to functional-dimension data exclusively. We transformed the absolute ratings into rankings using the rankdata library [113]. We employed rank as the dependent variable while maintaining all the steps outlined in the previous research questions. We applied various machine learning models, including CBR, RF, BR, and XGB, to all five functional-dimension apps and assessed the performance consistency using MSE, MAE, and RMSE metrics. Additionally, we evaluated the Normalized RMSE scores based on the standard deviation and interquartile range of data to compare different apps at the functional dimension 5.3.1. After determining the best model for each functional-dimension app, we evaluated the feature importance scores. The

applications were then grouped and ranked based on these scores using the Scott-Knott ESD 5.3.2. Upon obtaining the ranking of features for each functional dimension, we compared the features identified through the relative ranking mechanism with those derived from absolute ratings. This comparison enabled us to determine distinct prioritization strategies among competitors and analyze functionality-specific features.

Table 5.6 Model performance on relative ranking in functional dimension

Dimension		Model	MSE	MAE	RMSE	NRMSE-SD	NRMSE-IQR
Functional	Fitness-Training	Case Based Reasoning	0.98	0.78	0.99	1.58	1.2
		Bagging	0.99	0.79	0.99	1.59	1.22
		Random Forest	0.78	0.7	0.88	1.41	1.08
		XG Boost	1.9	1.14	1.4	2.24	1.71
		Dummy	2.6	1.32	1.62	2.5	1.9
Functional	Football-Betting	Case Based Reasoning	1.46	1.04	1.21	2.41	2.33
		Bagging	1.62	1.06	1.27	2.54	2.45
		Random Forest	1.41	1	1.19	2.37	2.29
		XG Boost	2.21	1.22	1.48	2.9	2.8
		Dummy	2.5	1.8	1.5	3.18	3.08
Functional	Football-Streaming	Case Based Reasoning	1.4	0.99	1.18	1.9	1.15
		Bagging	1.44	1	1.2	1.9	1.57
		Random Forest	1.2	0.95	1.11	1.79	1.47
		XG Boost	2.21	1.21	1.49	2.38	1.95
		Dummy	3.01	1.4	1.7	2.79	2.28
Functional	Fitness-Tracking	Case Based Reasoning	1.32	0.96	1.14	2.16	1.82
		Bagging	1.25	0.92	1.12	2.11	1.78
		Random Forest	1.11	0.87	1.05	1.99	1.67
		XG Boost	2.15	1.19	1.46	2.76	2.33
		Dummy	2.79	1.38	1.67	3.14	2.65
Functional	Football-News	Case Based Reasoning	1.29	0.97	1.13	2.09	1.92
		Bagging	1.53	1.04	1.23	2.27	2.09
		Random Forest	1.4	1.03	1.18	2.17	2
		XG Boost	2.27	1.23	1.5	2.77	2.55
		Dummy	3.06	1.44	1.74	3.21	2.96

Results

Random Forest Regressor consistently performed well across all functional dimensions, with relatively low MSE, MAE, and RMSE values. Models keep a similar trend to absolute rating prediction. Table 5.6 shows the performance of all 4 regression models to the baseline dummy regressor. For all 5 functional-dimension apps, regression models perform better than the baseline dummy regressor, and Random Forest is consistently the best-performing model across all the dimensions. Fitness-Training Apps and Fitness-Tracking Apps demonstrated similar performance patterns across different models, suggesting a possible correlation between the two functional dimensions. The performance differences between models across functional dimensions highlight the need for a customized approach to modelling in each functional dimension to obtain the best results.

Table 5.7 Feature importance among competitors based on relative ranking

Level		Features	Rank
Functional	Fitness-Training	Length of Description	1
		Screenshots	2
		Updated, Reviews	3
		Length of Name	4
		Genre, prevent device from sleeping	5
Functional	Football-Betting	Updated	1
		Screenshots	2
		Length of Description, Reviews	3
		Ratings, offers IAP	4
		Content Rating	5
Functional	Football-Streaming	Reviews	1
		Length of Name, Length of Description	2
		Updated	3
		Genre, Screenshots	4
		Released, allow Wi-Fi Multicast reception, change network connectivity, use accounts on the device, change your audio settings, draw over other apps	5
Functional	Fitness-Tracking	Screenshots	1
		Reviews	2
		Updated, Length of Description	3
		take pictures and videos, content Rating, run at - startup, view wifi connections, Length of Name	4
		receive data from Internet, precise - location (GPS and network-based), videoImage, control vibrations, offersIAP, use accounts on the device	5
Functional	Football-News	Lenth of Name	1
		Screenshots	2
		Reviews	3
		precise location (GPS and network-based), Length of Description	4
		Content Rating	5

Note: Each rank contains features that are statistically indistinguishable from each other in terms of their importance score.

For highly competitive apps categories, more important features were identified in the relative ranking predictions such as for Fitness Tracking new model identified 8 new features such as *take pictures and videos, run at startup, view wifi connections, receive data from Internet, precise location (GPS and network-based), videoImage, control vibrations, use accounts on the device*. The analysis based on the relative ranking has revealed more features for functional dimensions characterized by a highly competitive landscape,

such as Fitness-Tracking and Fitness-Streaming. Fig 5.3 demonstrates that Fitness Training, Fitness Tracking, and Football Streaming applications exhibit highly competitive landscapes. No unique features were identified for Fitness-Tracking and Football-Streaming, as the apps appear to be clustered regarding quality and user satisfaction (Table 5.5). The relative ranking-based feature importance revealed Fitness-Tracking specific features such as *videoImage*, *control vibrations*, and *use accounts on the device*. App-specific features, including *prevent device from sleeping* for Fitness-Training Apps and *precise location (GPS and network-based)* for Fitness-Tracking and Football-News Apps, are crucial within their respective functional dimensions. *Content Rating* emerges as a significant feature for Football-Betting and Football-News Apps, implying that age-appropriateness may be a consideration for users within these categories.

Offers IAP (In-App Purchases) is an essential feature for Football-Betting Apps, suggesting that monetization strategies contribute to the success of apps in this category. The analysis among competitors assists in narrowing down specific features related to user experience. For instance, Football-Streaming features such as *change network connectivity*, *use accounts on the device*, *change your audio settings*, and *draw over other apps* highlight the importance of user experience while watching videos.

Certain features are unique to specific functional dimensions, emphasizing the necessity for tailored strategies and insights when optimizing sports apps across different categories. As a result, the null hypothesis for the research question, H_0 : *There is no significant relationship between the features of an app and its ranking among competitors*, is rejected.

Analyzing applications at the functional dimension provides more targeted insights compared to categorical or market-dimension analysis. To comprehend the features influencing competitors and the competitive landscape, employing a relative ranking-based approach can facilitate the identification of competitor-specific information, particularly in highly competitive environments.

5.4 Discussions

In this section, we discuss the implications of our results.

The heterogeneity of information across distinct app dimensions underscores the significance of addressing the subtleties and nuances inherent in the data when creating machine learning models tailored to different app dimensions: Our investigation in RQ1 reveals that regression models exhibit superior performance at functional

dimensions compared to the broader market dimension. However, it is crucial to note that our study's scope is limited to 1,457 sports apps, while the functional apps are in the hundreds. In comparison, there are 47,000 sports apps available on the Google Play Store[1]. As the dataset increases, models are likely to become more complex, leading to potential challenges in effectively capturing the unique characteristics and competitive landscapes of specific app functionalities within sports apps and their functional dimensions. This complexity is further highlighted by the divergent competition landscapes observed in Fig 5.3. Consequently, it is essential to consider these factors when developing models that can accurately reflect the diverse needs and intricacies inherent in sports apps across different dimensions and functionalities.

Using explainable machine learning models on multidimension analysis helps to identify the feature importance that can aid the App store optimization factors:

The use of explainable machine learning models, particularly in multidimension analysis, can assist in identifying the critical features that can aid in optimizing an app's visibility in the app store. App store optimization services that cater to specific functional dimensions often provide paid services to enhance an application's reach, indicating the significance of identifying the most crucial information to improve an app's visibility. Our study's results in RQ2 suggest that a simple multidimension analysis can help app developers determine the most specific information for optimizing their apps, which can significantly contribute to app store optimization factors. The multidimension analysis approach allows for the identification of the unique characteristics and competitive landscapes of specific app functionalities within sports apps and their functional dimensions, despite the dataset's increased complexity as the number of apps grows.

A combination of absolute rating-based prediction and relative ranking-based prediction give different aspect of competitors:

In highly competitive app markets, app ratings can be very similar, making it difficult to differentiate between them based on their absolute ratings alone. In such cases, it's important to identify the small details that can aid an app's success. That's where relative ranking-based prediction comes in. By analyzing an app's relative ranking within its category, we can identify specific information that may be omitted in absolute rating-based predictions. Therefore, using both absolute rating-based and relative ranking-based predictions can provide a more comprehensive understanding of an app's competitive landscape, which is especially useful in highly competitive scenarios. Our study's findings from RQ2 and RQ3 suggest that, for functional dimensions, apps should focus more on the functionality-based permissions of their apps to improve their ratings. This finding highlights the importance of considering app functionalities when developing machine learning models for predicting app ratings. It also emphasizes the need to use a

combination of absolute rating-based and relative ranking-based predictions to gain a deeper understanding of an app’s competitive landscape, especially for specific functionalities within the application.

5.5 Threats to validity

This section discusses threats to the validity of our study.

5.5.1 Internal Validity

The primary objective of our study is to comprehend user rating patterns within various app categories and functional dimensions. Prior research has indicated inconsistencies between app ratings and sentiment analysis of corresponding reviews, suggesting that ratings may not always accurately reflect user opinions [96]. Nonetheless, we assume that user-provided ratings are a genuine representation of their intended assessment of the applications. It is essential to consider potential biases in these ratings, which may arise from socio-political factors, such as rating bombardments, where groups of users intentionally give low ratings to affect an app’s overall score [112]. Our study operates under the assumption that the Google Play Store has appropriately addressed and mitigated these biases during data collection, ensuring that the ratings analyzed in our research are an accurate representation of users’ experiences and opinions.

5.5.2 External Validity

In this study, we analyze more than 12,000 apps spanning various functional dimensions. The smallest functional dimension consists of 181 applications, which is a sufficient sample size for conducting regression analysis. However, it is important to acknowledge that this sample may not capture the entire competitive landscape within each functional dimension. To assess the reliability of our results, we compared five distinct groups of competitors within the sports app category.

The primary objective of this paper is to demonstrate the importance of multi-dimension analysis in understanding user rating patterns and feature importance across different functional dimensions. By identifying influential factors at various dimensions, we can apply our approach to additional functionalities, thus uncovering more functional-dimension insights. Expanding our approach to encompass a broader range of app functionalities will not only maintain the motivation of our current study but also enhance its generalizability.

It is essential to note that our data is limited to the Google Play Store, one of the largest app marketplaces. We conducted a comparison of the features available on F-Droid and the Apple App Store and found them to be comparable, suggesting that our findings may be applicable to other platforms. To ensure transparency and replicability, we have meticulously detailed our data preprocessing and cleaning steps. Furthermore, we plan to provide a replication package for our research, enabling other researchers to replicate our project using data from different app stores. This will help validate the robustness of our findings across various platforms and contexts, enhancing the external validity of our study.

5.5.3 Construct Validity

In RQ1 and RQ3, we employ Case-based reasoning, for which we selected the optimal number of neighbours (K) by minimizing the mean and median error scores for K ranging from 1 to 20. It is important to note that the optimal K value is subjective for each category, as evidenced by our results and previous research [9].

To demonstrate the significance of multi-dimension analysis, we needed to classify the applications into functional dimensions. We achieved this classification using a combination of keyword-based analysis of app descriptions and manual assessment to further refine the categorization. We examined the classified sports types and functionalities and analyzed the descriptions of 1,457 sports apps. This approach, however, may introduce potential biases based on the perceived functionality of the apps. Nevertheless, we expect developers to explicitly mention the app’s functionalities in the description to facilitate user understanding.

By elaborating on our classification process, we aim to ensure the construct validity of our study. This includes acknowledging the potential biases that might arise from our classification method and taking necessary steps to mitigate them. It is crucial to continuously refine our classification approach to accurately capture the essence of the various functional dimensions within the sports app ecosystem. A robust classification will ultimately enhance the validity of our findings and contribute to a more comprehensive understanding of the factors influencing user ratings across different functional dimensions.

5.6 Chapter summary

This study conducts a comprehensive, multidimension analysis to examine rating patterns across over 12,000 applications. The data is organized into three distinct dimensions: market dimension, categorical dimension, and functional dimension. Our findings suggest that regression models exhibit enhanced performance at lower dimensions, as the granularity of in-

formation is more substantial compared to higher dimensions. The random forest regression model consistently outperforms other models across all dimensions and is employed to identify features influencing rating patterns. We adapt the feature importance score to feature ranking utilizing the Scott-Knott Effect Size Difference (ESD) method to discern the most impactful ranks. We identified the most influencing features regardless of the dimensions, such as reviews, screenshots, and description length. Furthermore, this analysis helps pinpoint category-specific features, such as Bluetooth connection permissions for sports apps and wallpaper settings access for personalized apps. At the functional dimension, we identified specific features that are affecting the rating among competitors (e.g., network connection permissions, including change/intercept network settings and Wi-Fi connectivity, were found to be crucial for football betting apps). Due to the highly competitive landscape among competitors, the relative ranking-based approach proved effective in accurately identifying more than five functional-dimension features (e.g., audio settings adjustments and overlays for football streaming and device account usage and Bluetooth settings for fitness tracking). Despite the dynamic nature of the app store and the rapid growth of the sports app market, this research presents a wealth of opportunities for developers to refine their applications and make informed business decisions. We encourage future research to expand upon this work by examining additional categories to understand better the specific features that contribute to the success of individual app categories and provide more meaningful insights and recommendations.

CHAPTER 6 CONCLUSION

In this chapter, we summarise our results and conclude the thesis. Besides, we highlight the limitations of our approaches and outline some directions for future work.

6.1 Summary

As advancements in smartphone technology, 5G, and machine learning continue to accelerate, the mobile application industry has become an increasingly significant component of software development. Because of the improvement in mobile app stores, it is also important to note that user-perceived information has increased with very vital information related to the app quality and improvement. The Android app marketplace, Google Play Store, currently hosts approximately 2.6 million applications. User perception of the quality of the app provides direct insight into the app's performance in relation to the description and promises made on the app store. This data is crucial for developers to enhance the user experience, ensure user retention and expand the user base through recommendations. Lastly, positive user perception also improves app store rankings, increasing the application's visibility and reach in the competitive marketplace. Given the substantial user engagement associated with a single category (e.g., sports applications), it is imperative to scrutinize this particular category independently.

In this study, we investigate a large-scale multidimension (market dimension, category dimension and functional dimension) dataset of apps to analyze user-perceived data, specifically user reviews and ratings available on the Google Play Store, considering sports apps as a case study. We primarily analyze user reviews and ratings from the Google Play Store for sports apps to understand the topics that users talk about in their reviews and their primary concerns using sentiment analysis. We also compare the topics and sentiments discussed at category and functional dimensions with market-dimension studies conducted in previous research. Secondly, the study centers on comprehending the important features influencing the user ratings of mobile apps, detailed into three distinct dimensions. Moreover, we compare the differences in important features for each dimension and employ a methodology for identifying significant features based on relative ranking at the functional dimension, where the competition is fiercely competitive. This approach aids in identifying the important features that provide an edge to mobile apps amidst cut-throat competition.

In chapter 2, we present background knowledge of concepts and terminologies that are helpful

to understand our empirical studies better. Chapter 3 provided a comprehensive review of the literature related to various topics, including the analysis of general app reviews, sports app analysis, rating predictions for general apps, and previous research on multidimension app data analysis.

In chapter 4, we did an empirical study on large-scale sports apps where we examined the features of a substantial collection (2,058) of sports applications available on the Google Play store, as well as their corresponding user reviews (40,266). Where identified a diverse range of sports applications, encompassing 16 sports varieties (e.g., Football, Cricket, Baseball) and 15 primary functions (e.g., Betting, Betting Tips, Training, Tracking). Utilizing automated topic modelling, we extracted 14 topics from user feedback, with 3 topics being exclusive to sports applications (*prediction accuracy*, *timeliness*, and *tool precision*). Sentiment analysis revealed that user grievances primarily stem from advertisements and various quality concerns (e.g., software bugs, quality of content, streaming quality) within sports applications. Our study shed light on a singular application category in our large-scale investigation; we demonstrate that a targeted analysis of specific app types (e.g., sports applications) yields more pertinent insights in comparison to a broader examination of applications across multiple categories while still maintaining relevance for a considerable number of applications.

In chapter 5, we conducted a multidimension analysis of over 12,000 applications to examine rating patterns. The data was organized into three dimensions: market, categorical, and functional. The study found that regression models perform better at lower dimensions and that the random forest regression model outperforms other models across all dimensions. The most influential features, such as reviews, screenshots, and description length, were identified. Through a multidimension analysis, it was possible to identify features that are unique to categorical dimensions (e.g., the requirement of Bluetooth connection permissions for sports applications) and those that are significant at the functional dimension (e.g., such as the criticality of network connection permissions for football betting applications) which may not be discernible through market-dimension analysis. The study proves that more intricate details have been identified in a multidimension approach, and the market dimension dataset results are not enough for the developers to refine their applications and make informed business decisions. Our study is the first to analyze a sports app category on a large scale and compare the importance of multidimension analysis for reviews and ratings. These results can help the developer to tailor their application specific to their functionalities to improve the quality of the apps. We hope that our research will advocate for subsequent research to conduct studies to understand the importance of user-perceived metrics as well as more in-depth studies on specialized app categories, to produce more significant observations and recommendations tailored to those specific categories.

Challenges: The first challenge we faced was the limited availability of data, despite the existence of previous studies in similar areas. Replicating these studies proved challenging due to missing information in algorithms or the unavailability of data. This required us to carefully select and curate the data sources we used and also to develop new methods for extracting and analyzing data.

The second challenge we encountered was the fast-paced and dynamic nature of the sports app market. This meant that application data was changing on a day-to-day basis and that the data we crawled required continuous updates. Additionally, Google Play Store metrics add and remove applications, which further complicates the process of data collection and analysis.

A final challenge we faced was manual labelling, which was necessary to determine the specific functionalities of sports apps and the algorithms being used. The genres provided by the Google Play Store were often highly generalized, requiring us to conduct manual analysis to identify the specific features of the apps. While we initially adopted a keyword-based approach to labelling, we found that many applications had significant overlap in keywords, requiring us to use a combination of manual and automated labelling approaches.

Overall, these challenges required us to be creative and adaptable in our approach to data collection and analysis and to develop new methods for addressing the limitations of the available data. By doing so, we were able to conduct a rigorous and comprehensive analysis of sports apps and their functionalities, as well as gain new insights into the sports app market.

6.2 Limitations of the proposed approaches

1. In our study we used one of the biggest app distribution system, Google Play Store; however, it can be generalized across the other application store. The developers have to optimize the code based on the features they have.
2. Google Play Store undergoes lot of changes on the data that they provide in terms of UI changes which makes availability or unavailability of data that is available at the moment. Google Play Store also optimizes the reviews and the application based on their internal guidelines thus the models that we are building will be biased on the approach they have adopted or data could be biased with user review bombing [101, 112].
3. The data available in the Google Play Store for a specific app is a very small represen-

tation of the app and its performance in general. So limiting the rating prediction only based on the metadata available on the Google Play Store might not give the complete analysis.

6.3 Future research

Our research aims to expand in four directions. First, we intend to develop a system that can be applied to data stores beyond the Google Play Store, such as the Apple App Store and Samsung Store, with the goal of creating a comprehensive open-sourced contribution that can facilitate app store optimization. Second, our analysis of the top functionalities of sports apps revealed that football betting is a popular category, with many apps employing various types of analytical and predictive models. By examining the usage of machine learning models in mobile application code, we can improve the success of sports apps. Third, we observed that many app permissions have a significant relationship to the app rating at both categorical and functional levels. Analyzing the privacy and security concerns and issues that users face and their impact on the app rating can help inform data privacy policies. Finally, we plan to understand user engagement with apps, including the time that users spend on the application, the number of active users, and other relevant metrics. By examining user engagement, we can gain a detailed understanding of user ratings and review sentiment.

REFERENCES

- [1] Appbrain, “Google Play category statistics table,” <https://www.appbrain.com/stats/android-market-app-categories>, Accessed 2022-05-01.
- [2] A. Følstad, “Users’ design feedback in usability evaluation: a literature review,” *Human-centric Computing and Information Sciences*, vol. 7, no. 1, p. 19, 2017.
- [3] K. Kalaichelavan, H. Malik, N. Husnu, and S. Sreenath, “What do people complain about drone apps? a large-scale empirical study of google play store reviews,” *Procedia Computer Science*, vol. 170, pp. 547–554, 2020.
- [4] H. Khalid, E. Shihab, M. Nagappan, and A. E. Hassan, “What do mobile app users complain about?” *IEEE software*, vol. 32, no. 3, pp. 70–77, 2014.
- [5] M. Karagkiozidou, C. Ziakis, M. Vlachopoulou, and T. Kyrkoudis, “App store optimization factors for effective mobile app ranking,” in *Strategic Innovative Marketing and Tourism: 7th ICSIMAT, Athenian Riviera, Greece, 2018*. Springer, 2019, pp. 479–486.
- [6] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh, “Why people hate your app: Making sense of user feedback in a mobile app store,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1276–1284.
- [7] 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)*. IEEE, 2017, pp. 55–58.
- [8] E. Guzman and W. Maalej, “How do users like this feature? a fine grained sentiment analysis of app reviews,” in *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 2014, pp. 153–162.
- [9] F. Sarro, M. Harman, Y. Jia, and Y. Zhang, “Customer rating reactions can be predicted purely using app features,” in *2018 IEEE 26th International Requirements Engineering Conference (RE)*. IEEE, 2018, pp. 76–87.
- [10] A. Mahmood, “Identifying the influence of various factor of apps on google play apps ratings,” *Journal of Data, Information and Management*, vol. 2, pp. 15–23, 2020.

- [11] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Predicting numeric ratings for google apps using text features and ensemble learning," *ETRI Journal*, vol. 43, no. 1, pp. 95–108, 2021.
- [12] Statista, "Time spent per day with mobile apps worldwide as of q4 2020, by category," <https://www.statista.com/statistics/1272800/worldwide-mobile-apps-time-spent-daily/>, 2021, [Accessed: 26-Mar-2023].
- [13] R. Vasa, L. Hoon, K. Mouzakis, and A. Noguchi, "A preliminary analysis of mobile app user reviews," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, ser. OzCHI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 241–244. [Online]. Available: <https://doi.org/10.1145/2414536.2414577>
- [14] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, vol. 43, no. 3, pp. 345–354, 2006.
- [15] "Data ai," <https://www.data.ai/en/>, accessed on March 26, 2023.
- [16] "Sensor tower," <https://sensortower.com/>, accessed on March 26, 2023.
- [17] MobileAction, "App store optimization (aso): The main factors of aso," <https://www.mobileaction.co/app-store-optimization-aso/#the-main-factors-of-aso>, accessed March 08 2023.
- [18] "Appbot," <https://appbot.co/>, accessed on: March 27, 2023.
- [19] "Firebase crashlytics," <https://firebase.google.com/products/crashlytics>, accessed on: March 27, 2023.
- [20] "Mixpanel," <https://mixpanel.com/>, accessed on: March 27, 2023.
- [21] D. Huang and H. Wu, *Mobile cloud computing: foundations and service models*. Morgan Kaufmann, 2017.
- [22] Statista. (2021) Global market share held by smartphone operating systems. [Online]. Available: <https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/>
- [23] —, "Number of apps available in leading app stores," <https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/>, 2023, accessed on March 27, 2023.

- [24] Google, “Google play distribution,” <https://developer.android.com/distribute>, [Accessed March 27, 2023].
- [25] Android, “Android security and privacy,” <https://www.android.com/safety/>, [Accessed March 27, 2023].
- [26] S. Hassan, H. Li, and A. E. Hassan, “On the importance of performing app analysis within peer groups,” *arXiv preprint arXiv:2204.03794*, 2022.
- [27] S. Keertipati, B. T. R. Savarimuthu, and S. Licorish, “Approaches for prioritizing feature improvements extracted from app reviews,” 06 2016, pp. 1–6.
- [28] D. Sheen, “Evolution of analytical data in sports,” *Medium*, December 2020. [Online]. Available: <https://dabidsheen.medium.com/evolution-of-analytical-data-in-sports-8a096ace464c>
- [29] A. Hutchinson, “Reddit shares new insights into nfl engagement in the app,” *Social Media Today*, September 2021. [Online]. Available: <https://www.socialmediatoday.com/news/reddit-shares-new-insights-into-nfl-engagement-in-the-app-infographic/629049/>
- [30] C. Gao, J. Zeng, M. R. Lyu, and I. King, “Online app review analysis for identifying emerging issues,” in *Proceedings of the 40th International Conference on Software Engineering*, 2018, pp. 48–58.
- [31] C. Tao, H. Guo, and Z. Huang, “Identifying security issues for mobile applications based on user review summarization,” *Information and Software Technology*, vol. 122, p. 106290, 2020.
- [32] W. Luiz, F. Viegas, R. Alencar, F. Mourão, T. Salles, D. Carvalho, M. A. Gonçalves, and L. Rocha, “A feature-oriented sentiment rating for mobile app reviews,” in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1909–1918.
- [33] S. McIlroy, N. Ali, H. Khalid, and A. E Hassan, “Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews,” *Empirical Software Engineering*, vol. 21, no. 3, pp. 1067–1106, 2016.
- [34] A. Di Sorbo, G. Grano, C. Aaron Visaggio, and S. Panichella, “Investigating the criticality of user-reported issues through their relations with app rating,” *Journal of Software: Evolution and Process*, vol. 33, no. 3, p. e2316, 2021.

- [35] 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*, 2014, pp. 767–778.
- [36] M. Wano and J. Iio, “Relationship between reviews at app store and the categories for software,” in *2014 17th International Conference on Network-Based Information Systems*, 2014, pp. 580–583.
- [37] J. J. Aman, J. Smith-Colin, and W. Zhang, “Listen to e-scooter riders: Mining rider satisfaction factors from app store reviews,” *Transportation Research Part D: Transport and Environment*, vol. 95, p. 102856, 2021.
- [38] M. Fazzini *et al.*, “Characterizing human aspects in reviews of covid-19 apps,” in *Proceedings of the 9th IEEE/ACM International Conference on Mobile Software Engineering and Systems*. IEEE/ACM, 2022.
- [39] Z. S. H. Abad, S. D. V. Sims, A. Cheema, M. B. Nasir, and P. Harisinghani, “Learn more, pay less! lessons learned from applying the wizard-of-oz technique for exploring mobile app requirements,” in *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2017, pp. 132–138.
- [40] M. van Vliet, E. C. Groen, F. Dalpiaz, and S. Brinkkemper, “Identifying and classifying user requirements in online feedback via crowdsourcing,” in *Requirements Engineering: Foundation for Software Quality - 26th International Working Conference, REFSQ 2020, Pisa, Italy, March 24-27, 2020, Proceedings [REFSQ 2020 was postponed]*, ser. Lecture Notes in Computer Science, N. H. Madhavji, L. Pasquale, A. Ferrari, and S. Gnesi, Eds., vol. 12045. Springer, 2020, pp. 143–159.
- [41] M. V. Phong, T. T. Nguyen, H. V. Pham, and T. T. Nguyen, “Mining user opinions in mobile app reviews: A keyword-based approach (t),” in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2015, pp. 749–759.
- [42]
- [43] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, “Analysing app reviews for software engineering: a systematic literature review,” *Empirical Software Engineering*, vol. 27, no. 2, p. 43, 2022.

- [44] M. Zečević, D. Mijatović, M. K. Koklič, V. Žabkar, and P. Gidaković, “User perspectives of diet-tracking apps: Reviews content analysis and topic modeling,” *Journal of Medical Internet Research*, vol. 23, no. 4, p. e25160, 2021.
- [45] R. Albalawi, T. H. Yeap, and M. Benyoucef, “Using topic modeling methods for short-text data: A comparative analysis,” *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020.
- [46] K. Srisopha, C. Phonsom, M. Li, D. Link, and B. Boehm, “On building an automatic identification of country-specific feature requests in mobile app reviews: Possibilities and challenges,” in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops, ICSEW’20*. Association for Computing Machinery, 2020, pp. 494–498.
- [47] E. Guzman and W. Maalej, “How do users like this feature? a fine grained sentiment analysis of app reviews,” in *2014 IEEE 22nd international requirements engineering conference (RE)*. Ieee, 2014, pp. 153–162.
- [48] N. Genc-Nayebi and A. Abran, “A systematic literature review: Opinion mining studies from mobile app store user reviews,” *Journal of Systems and Software*, vol. 125, pp. 207–219, Mar 2017.
- [49] X. Gu and S. Kim, “" what parts of your apps are loved by users?"(t),” in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2015, pp. 760–770.
- [50] A. Finkelstein, M. Harman, Y. Jia, F. Sarro, and Y. Zhang, “Mining app stores: Extracting technical, business and customer rating information for analysis and prediction,” *RN*, vol. 13, p. 21, 2013.
- [51] A. Finkelstein, M. Harman, Y. Jia, W. Martin, F. Sarro, and Y. Zhang, “App store analysis: Mining app stores for relationships between customer, business and technical characteristics,” *RN*, vol. 14, no. 10, pp. 1–24, 2014.
- [52] S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, “Discrepancy detection between actual user reviews and numeric ratings of google app store using deep learning,” *Expert Systems with Applications*, vol. 181, p. 115111, 2021.
- [53] M. Harman, Y. Jia, and Y. Zhang, “App store mining and analysis: Msr for app stores,” in *2012 9th IEEE working conference on mining software repositories (MSR)*. IEEE, 2012, pp. 108–111.

- [54] A. Finkelstein, M. Harman, Y. Jia, W. Martin, F. Sarro, and Y. Zhang, “Investigating the relationship between price, rating, and popularity in the blackberry world app store,” *Information and Software Technology*, vol. 87, pp. 119–139, 2017.
- [55] D. Erić, R. Bačik, and I. Fedorko, “Rating decision analysis based on ios app store data,” *Quality Innovation Prosperity*, vol. 18, no. 2, pp. 27–37, 2014.
- [56] N. Gnotthivongsa and D. J. Huang, “Rating prediction for mobile applications via collective matrix factorization considering app categories,” in *Journal of Physics: Conference Series*, vol. 1993, no. 1. IOP Publishing, 2021, p. 012034.
- [57] I. J. Mojica Ruiz, M. Nagappan, B. Adams, T. Berger, S. Dienst, and A. E. Hassan, “Examining the rating system used in mobile-app stores,” *IEEE Software*, vol. 33, no. 6, pp. 86–92, 2016.
- [58] Google Play, “Compare your app’s android vitals and ratings with custom peer groups,” Google Play Support, 2021. [Online]. Available: <https://support.google.com/googleplay/android-developer/answer/9842755?hl=en>
- [59] F. Dalpiaz and M. Parente, “Re-swtot: from user feedback to requirements via competitor analysis,” in *Proceedings of the 25th International Working Conference on Requirements Engineering: Foundation for Software Quality*, ser. REFSQ ’19, 2019, pp. 55–70.
- [60] M. Assi, S. Hassan, Y. Tian, and Y. Zou, “Featcompare: Feature comparison for competing mobile apps leveraging user reviews,” *Empirical Software Engineering*, vol. 26, no. 5, pp. 1–38, 2021.
- [61] J. Dallinga, M. Janssen, J. Van Der Werf, R. Walravens, S. Vos, M. Deutekom *et al.*, “Analysis of the features important for the effectiveness of physical activity-related apps for recreational sports: Expert panel approach,” *JMIR mHealth and uHealth*, vol. 6, no. 6, p. e9459, 2018.
- [62] F. Modave, J. Bian, T. Leavitt, J. Bromwell, C. Harris III, and H. Vincent, “Low quality of free coaching apps with respect to the american college of sports medicine guidelines: a review of current mobile apps,” *JMIR mHealth and uHealth*, vol. 3, no. 3, p. e4669, 2015.
- [63] P. Shen *et al.*, “Analysis of user needs of sports and fitness apps from the perspective of “healthy china”,” *Academic Journal of Business & Management*, vol. 2, no. 5, 2020.

- [64] S. Paganini, Y. Terhorst, L. B. Sander, S. Catic, S. Balci, A.-M. Küchler, D. Schultchen, K. Plaumann, S. Sturmbauer, L. V. Krämer *et al.*, “Quality of physical activity apps: systematic search in app stores and content analysis,” *JMIR mHealth and uHealth*, vol. 9, no. 6, p. e22587, 2021.
- [65] A. Muntaner-Mas, A. Martinez-Nicolas, C. J. Lavie, S. N. Blair, R. Ross, R. Arena, and F. B. Ortega, “A systematic review of fitness apps and their potential clinical and sports utility for objective and remote assessment of cardiorespiratory fitness,” *Sports Medicine*, vol. 49, no. 4, pp. 587–600, 2019.
- [66] W. Martin, F. Sarro, Y. Jia, Y. Zhang, and M. Harman, “A survey of app store analysis for software engineering,” *IEEE Transactions on Software Engineering*, vol. 43, no. 9, pp. 817–847, 2017.
- [67] Q. Chen, C. Chen, S. Hassan, Z. Xing, X. Xia, and A. E. Hassan, “How should i improve the UI of my app? a study of user reviews of popular apps in the Google Play,” *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 30, no. 3, pp. 1–38, 2021.
- [68] K. Frie, J. Hartmann-Boyce, S. Jebb, C. Albury, R. Nourse, P. Aveyard *et al.*, “Insights from google play store user reviews for the development of weight loss apps: mixed-method analysis,” *JMIR mHealth and uHealth*, vol. 5, no. 12, p. e8791, 2017.
- [69] H. M. Brown, T. Bucher, C. E. Collins, and M. E. Rollo, “A review of pregnancy apps freely available in the google play store,” *Health Promotion Journal of Australia*, vol. 31, no. 3, pp. 340–342, 2020.
- [70] Grand View Research, “Sports Technology Market Size, Share Trends Analysis Report By Technology (Devices, Smart Stadium), By Sports (Basketball, Soccer, Tennis, Golf), By Region (APAC, Europe), And Segment Forecasts, 2021 - 2028,” <https://www.grandviewresearch.com/industry-analysis/sports-technology-market>, Accessed 2022-05-01.
- [71] Appbrain, “Google Play Ranking: The Top Free Sports Apps in the United States,” https://www.appbrain.com/stats/google-play-rankings/top_free/sports/us, Accessed 2022-05-01.
- [72] I. Malavolta, S. Ruberto, T. Soru, and V. Terragni, “End users’ perception of hybrid mobile apps in the google play store,” in *2015 IEEE International Conference on Mobile Services*. IEEE, 2015, pp. 25–32.

- [73] S. Hassan, C.-P. Bezemer, and A. E. Hassan, "Studying bad updates of top free-to-download apps in the google play store," *IEEE Transactions on Software Engineering*, vol. 46, no. 7, pp. 773–793, 2018.
- [74] A. A. Al-Subaih, F. Sarro, S. Black, L. Capra, M. Harman, Y. Jia, and Y. Zhang, "Clustering mobile apps based on mined textual features," in *Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement*, 2016, pp. 1–10.
- [75] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "Yake! collection-independent automatic keyword extractor," in *European Conference on Information Retrieval*. Springer, 2018, pp. 806–810.
- [76] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [77] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [78] E. Morgulev, O. H. Azar, and R. Lidor, "Sports analytics and the big-data era," *International Journal of Data Science and Analytics*, vol. 5, no. 4, pp. 213–222, 2018.
- [79] R. Clarke, "Guidelines for the responsible application of data analytics," *Computer law & security review*, vol. 34, no. 3, pp. 467–476, 2018.
- [80] M. A. Hadi and F. H. Fard, "Aobtm: Adaptive online biterm topic modeling for version sensitive short-texts analysis," in *2020 IEEE international conference on software maintenance and evolution (ICSME)*. IEEE, 2020, pp. 593–604.
- [81] I. Akef, J. S. M. Arango, and X. Xu, "Mallet vs gensim: Topic modeling for 20 news groups report," *Univ. Ark. Little Rock Law J.*, [http://dx. doi. org/10.13140/RG](http://dx.doi.org/10.13140/RG), vol. 2, no. 19179, p. 39205, 2016.
- [82] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.
- [83] C. Rosen and E. Shihab, "What are mobile developers asking about? a large scale study using stack overflow," *Empirical Softw. Engg.*, vol. 21, no. 3, p. 1192–1223, jun 2016. [Online]. Available: <https://doi.org/10.1007/s10664-015-9379-3>

- [84] L. Cen, L. Si, N. Li, and H. Jin, “User comment analysis for android apps and csapi detection with comment expansion,” in *PIR@ SIGIR*, 2014.
- [85] L. Villarroel, G. Bavota, B. Russo, R. Oliveto, and M. Di Penta, “Release planning of mobile apps based on user reviews,” in *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*. IEEE, 2016, pp. 14–24.
- [86] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.
- [87] S. Elbagir and J. Yang, “Twitter sentiment analysis using natural language toolkit and vader sentiment,” in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 122, 2019, p. 16.
- [88] A. Borg and M. Boldt, “Using vader sentiment and svm for predicting customer response sentiment,” *Expert Systems with Applications*, vol. 162, p. 113746, 2020.
- [89] H. Akoglu, “User’s guide to correlation coefficients,” *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.
- [90] 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.
- [91] J. Schobel, M. Schickler, R. Pryss, H. Nienhaus, and M. Reichert, “Using vital sensors in mobile healthcare business applications: Challenges, examples, lessons learned,” in *International Conference on Web Information Systems and Technologies*, 2013, pp. 509–518.
- [92] StraitsResearch, “Mobile app development market size, share, growth, trends, analysis, report,” <https://straitsresearch.com/report/mobile-app-development-market>, 2023, accessed on February 8, 2023.
- [93] MobileAction, “How to combine ua & aso for mobile growth,” <https://www.mobileaction.co/blog/user-acquisition/how-to-combine-ua-aso-for-mobile-growth/>, 2021, accessed on February 8, 2023.
- [94] Business of Apps, “How ratings and reviews affect consumer decision to download apps,” <https://www.businessofapps.com/insights/>

- ratings-reviews-affect-consumer-decision-download-apps/, 2021, [Online; accessed 25-March-2023].
- [95] A. Strzelecki, “Application of developers’ and users’ dependent factors in app store optimization,” 2020.
 - [96] B. Chembakottu, H. Li, and F. Khomh, “A large-scale empirical study of android sports apps in the google play store,” *Dept. of Computer and Software Engineering, Polytechnique Montreal*, 2021.
 - [97] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, “An empirical comparison of model validation techniques for defect prediction models,” no. 1, 2017.
 - [98] S. McIlroy, N. Ali, and A. E. Hassan, “Fresh apps: an empirical study of frequently-updated mobile apps in the google play store,” *Empirical Software Engineering*, vol. 21, pp. 1346–1370, 2016.
 - [99] N. Viennot, E. Garcia, and J. Nieh, “A measurement study of google play,” in *The 2014 ACM international conference on Measurement and modeling of computer systems*, 2014, pp. 221–233.
 - [100] G. P. D. Help, “Why apps may be removed from google play,” https://support.google.com/googleplay/android-developer/answer/10532353?visit_id=638140250622748420-86029282&hl=en&rd=2, 2021, [Online; accessed 10-March-2023].
 - [101] F. Mohsen, D. Karastoyanova, and G. Azzopardi, “To remove or not remove mobile apps? a data-driven predictive model approach,” *arXiv preprint arXiv:2206.03905*, 2022.
 - [102] L. Myers and M. J. Sirois, “Spearman correlation coefficients, differences between,” *Encyclopedia of statistical sciences*, vol. 12, 2004.
 - [103] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, “An empirical study of the impact of modern code review practices on software quality,” *Empirical Software Engineering*, vol. 21, pp. 2146–2189, 2016.
 - [104] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book>
 - [105] P.-C. Chang, J.-J. Lin, and W.-Y. Dzan, “Forecasting of manufacturing cost in mobile phone products by case-based reasoning and artificial neural network models,” *Journal of Intelligent Manufacturing*, vol. 23, pp. 517–531, 2012.

- [106] W. Glodek and R. Harang, “Rapid permissions-based detection and analysis of mobile malware using random decision forests,” in *MILCOM 2013-2013 IEEE Military Communications Conference*. IEEE, 2013, pp. 980–985.
- [107] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
- [108] J. W. Osborne, *Regression & linear modeling: Best practices and modern methods*. Sage Publications, 2016.
- [109] J. Brownlee, “Regression metrics for machine learning,” <https://machinelearningmastery.com/regression-metrics-for-machine-learning/>, accessed on: March 25, 2023.
- [110] M. D. Science, “Normalizing the rmse,” <https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/>, January 2019, accessed on March 21, 2023.
- [111] X. Li, Y. Wang, S. Basu, K. Kumbier, and B. Yu, “A debiased mdi feature importance measure for random forests,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [112] S. Esports, “Google to help genshin impact recover from review bomb by deleting negative reviews on play store,” *Sportskeeda Esports*, October 2021. [Online]. Available: <https://www.sportskeeda.com/esports/google-help-genshin-impact-recover-review-bomb-deleting-negative-reviews-play-store>
- [113] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, and C. R. Harris, “scipy.stats.rankdata,” <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.rankdata.html>, July 2021, accessed on March 21, 2023.
- [114] V. Adams, *Metrics: What counts in global health*. Duke University Press, 2016.
- [115] S. A. Otto, M. Kadin, M. Casini, M. A. Torres, and T. Blenckner, “A quantitative framework for selecting and validating food web indicators,” *Ecological Indicators*, vol. 84, pp. 619–631, 2018.
- [116] L. Breiman, “Statistics department university of california berkeley, ca 94720. 2001,” *RANDOM FORESTS*.

- [117] “Add screenshots amp; videos to your listing,” Google Play Help, September 2021. [Online]. Available: <https://support.google.com/googleplay/android-developer/answer/9866151?hl=en#zippy=%2Cscreenshots>