TTAG+R: A Dataset of Google Play Store's Top Trending Android Games and User Reviews

Raheela Chand[1,*], Saif Ur Rehman Khan[1], Shahid Hussain[2], and Wenli Wang[2]
[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan
[2]Department of Computer Science and Software Engineering School of Engineering, Penn State University, USA
raheela.chand@yahoo.com
*corresponding author

Abstract—**Context: Android games are gaining wide attention from users in recent years. However, the existing literature reports alarming statistics about banning popular and top-trending Android apps. The popular gaming apps have been removed from Google Play Store due to various user concerns. Objectives: The goal of this work is twofold: (i) to assist the researchers and practitioners in identifying the state-of-the-art challenges, constraints, and compliments about Android apps for future Android-specific studies, and (ii) to encourage active users' perspectives on the Android development process because usability remains a core deciding factor about the success or failure of Android apps. Method: To accomplish this, we introduce a novel open-source dataset, Top Trending Android apps with their user Reviews (TTAG+R) in GitHub. Results and Contributions: Briefly, TTAG+R presents information about 245 top trending Android Free Games, 97 top trending Android Grossing Games, and 52 top trending Android Paid Games with a total of 8,423 user reviews in 12 different .csv files. The main contributions of this paper are: (i) provides one-place comprehensive data on Android Apps, (ii) describes various features of Android apps and their user reviews, (iii) reports the updated and latest knowledge about Android apps, and (iv) provides the data in an unfiltered form so that researchers may not find difficulty in using this dataset in their data-driven experimentation. From a research implication viewpoint, the dataset supports: (i) understanding the usability characteristics of Android apps, (ii) discovering current trends and pitfalls in Android apps, and (iii) analyzing the Android financial market. Conclusion and Future Work: Thus, TTAG+R is freely available to the research community, and useful for future enhancements in the Android domain. In the future, we plan to keep the data up-to-date with the most recent information for the continued usage of the dataset.**

Keywords—Google Play Store; Android Games; GitHub; Dataset; Repository; User Reviews

I. Introduction

The current software engineering research is motivated towards experimental studies by integrating Artificial Intelligence, Machine Learning, and various Data Mining techniques to automate the software development life cycle [1] [2] [3]. The data-driven empirical studies are more reliable and less prone to any external influence. Inspired by this, various datasets have been proposed by the research community. The availability of repositories has reduced biased approaches and increased authenticity in the results of experiments [4]. This paper introduces a new dataset, called Top Trending Android Apps and user Reviews (TTAG+R), to the research community. This dataset consists of a collection of .csv files that can be easily populated to identify various trends and targets in Android apps. The proposed dataset is a novel contribution for the reason that a few datasets have been proposed for analyzing the market value and the usage of Android apps. In 2017, Grano et al. [5] introduced a dataset of Android apps and user feedback. However, the authors have not updated their introduced dataset, thereby, lacking in accommodating the current knowledge. In contrast, Google Play Store apps is an updated Kaggle repository [6] but provides only statistical data on Android apps. Therefore, to the best of our knowledge, there is no repository that contains the current state-of-the-art data useful in investigating existing challenges in the Android app's context. In particular, this paper intends to encourage studies to actively incorporate users' apprehensions in the Android development process. This is due to the reason that the user is the main entity who will purchase and use the Android apps. Therefore, it is of vital importance to know about the customers' needs before providing a real-world solution. To acquire user insights, the proposed dataset is regarded as a one-place medium, which ignores the hectic investigating procedure for the collected user feedback that can be obtained through surveys or questionnaires. Generally speaking, the importance of user feedback is unquestionable that ultimately leads to the success of Android apps. The proper leverage of the proposed dataset can be beneficial in many ways. For example, Android apps require a novel specific list of checklists and testable quality metrics, which researchers can discover by manipulating the information provided in the TTAG+R. Likewise, Android apps are particularly vulnerable to security threats [7]. To address the above-mentioned issues, the developers need to be well aware of the user privacy concerns, which can be extracted from the user remarks (i.e. regarded as App reviews) provided in the dataset. Moreover, new revolutionary features are introduced in Android apps [8] on daily basis. Therefore, to compete in the business market, developers should be conscious of the aforementioned emerging trends. To handle this, the introduced dataset could be beneficial in discovering cutting-edge features. Moreover, some social factors can prevail in the market[9] that require the attention of developers while developing an Android app. In this regard, the proposed dataset provides useful information, which can be effective in dealing with the encountered social factors. Furthermore, keeping in mind the developer's perspective, we have also provided unfiltered information in TTAG+R with well-defined tables. Essentially, TTAG+R presents the data in five core components as the following:

2693-9371/22/$31.00 ©2022 IEEE
DOI 10.1109/QRS-C57518.2022.00093

1) The Catalogues contain Android games categories and divisions of the GPS.
2) The Game-Info files include publicly available information about Android games in Google Play Store (GPS).
3) The Lists of top charts.
4) The Reviews files contain the user's remarks on Android games.
5) The Super-set contains all the information provided in the dataset in one file.

To summarize contributions of this research paper are:

1) Provided a well-organized and single-comprehended repository of scattered information available on the internet for Android apps.
2) Extensively reviewed and parsed various features of Android Apps to follow multidimensional user usage of Android apps.
3) Gathered extensive and new knowledge about Android apps that can be used to understand current state-of-the-art challenges, trends, and user concerns.

In a nutshell, a developer must handle various challenges to develop a successful Android app. The challenges may vary from one Android app to another and from one person to other. To identify and overcome these challenges, TTAG+R assists by providing a plethora of information about Android Apps and their user remarks which would help in creating an architecture of the Android development process that can meet the company goals and satisfy user requirements.

The remaining part of this work is organized as follows: Section II describes the introduced dataset. The data collection methodology is provided in Section III. While Section IV provides future research opportunities. The encountered limitation and related work are mentioned in Sections V and VI, respectively. Finally, Section VII concludes the current work.

II. Dataset Description

This section describes the proposed dataset for the evaluation of Android gaming apps. As already mentioned, Top Trending Android Game and user Reviews (TTAG+R) is an open-source repository of GitHub. TTAG+R can be accessed through a link as mentioned in [10]. The following sections provide a detailed description, including context, execution framework, content, and format files, of the dataset:

A. Context of the Dataset

The information provided in the dataset is about top trending Android games and user reviews of the Google Play Store. The Google Play Store (GPS) also known as Google Play is the most popular and official publishing platform for Android apps. It provides around 2 million apps and games and generates over 120 billion dollars in earnings for developers [11]. However, a recent study shows that 30 percent of Android apps are banned due to poor quality [12]. Thus, it is a crucial challenge to logically identify malicious factors to help justify quality features in Android apps due to the non-availability of comprehensive knowledge. Motivated by this, we propose a novel dataset TTAG+R for the researchers and practitioners of the Android domain to perform future experiments on this dataset. Moreover, it could be useful in devising strategies to produce a high-quality Android app.

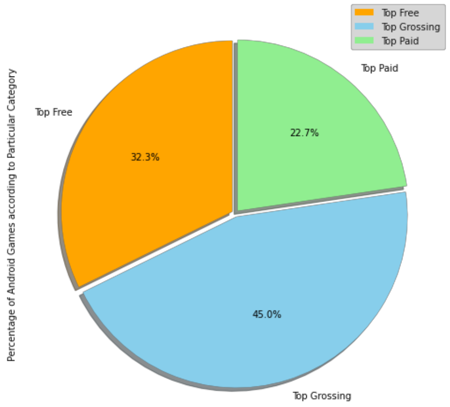


Figure 1. Percentage of Android Game Apps in Categories

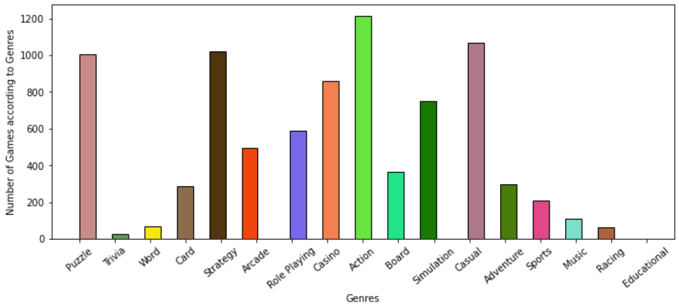


Figure 2. Distribution of Android Game Apps in Genres

B. Execution framework of the Dataset

The TTAG+R is stored in an open-source repository of GitHub. GitHub is a popular cloud-based service that helps developers to store and manage their codes. The link to access the dataset is given in the reference [10].

C. Content and files format

On the Google Play Store website, the Top Chart contains a list of trending Android apps. It is divided into three categories: (i) Top Paid, (ii) Top Free, and (iii) Top Grossing. Each category is further divided into various genres of Android apps. The information given in each category and genre is extracted as shown in Figures 1 and 2 and provided in .csv files in the dataset.

The contents of TTAG+R are described as follows:

1) The Game Info: Top Free Android Games file. It contains 2265 rows and 10 columns/fields of top free Android games i.e. Game ID, Name, URL, Ratings, Users, Price, Category, Genre, Date of Remarks, and Remarks.
2) The Game Info: Top Grossing Android Games file. It contains 3960 rows and 10 columns/fields of top free

Android games i.e. Game ID, Name, URL, Ratings, Users, Price, Category, Genre, Date of Remarks, and Remarks.

3) The Game Info: Top Paid Android Games file. It contains 2198 rows and 10 columns/fields of top free Android games i.e. Game ID, Name, URL, Ratings, Users, Price, Category, Genre, Date of Remarks, and Remarks.
4) Catalogue1: Android Games Genres file. It contains 17 rows and 2 columns/fields i.e. ID and Genre.
5) Catalogue2: Android Games Top Charts file. It contains 3 rows and 2 columns/fields i.e. ID and Category.
6) List1: Top Free Android Games file. It contains 96 rows and 7 columns/fields of top free Android games i.e. Game ID, Name, URL, Ratings, Users, Price, Category, and Genre.
7) List2: Top Grossing Android Games file. It contains 97 rows and 7 columns/fields of top free Android games i.e. Game ID, Name, URL, Ratings, Users, Price, Category, and Genre.
8) List3: Top Paid Android Games file. It contains 52 rows and 7 columns/fields of top free Android games i.e. Game ID, Name, URL, Ratings, Users, Price, Category, and Genre.
9) Reviews1: Top Free Android Games file. It contains 2265 rows and 8 columns/fields of top free Android games i.e. Review ID, Game Name, Ratings, Users, Category, Genre, Data of Remarks, and Remarks.
10) Reviews2: Top Grossing Android Games file. It contains 3960 rows and 8 col columns/fields of top Grossing Android games i.e. Review ID, Game Name, Ratings, Users, Category, Genre, Data of Remarks, and Remarks.
11) Reviews3: Top Paid Android Games file. It contains 2198 rows and 8 columns/fields of top Paid Android games i.e. Review ID, Game Name, Ratings, Users, Category, Genre, Data of Remarks, and Remarks.
12) Finally, The Super-set file. It contains 8423 rows and 10 columns.

III. Data Collection Methodology

In this section, the followed methodology to construct the dataset of top trending Android games and user reviews of the Google Play Store is described. The construction of TTAG+R is performed in 5 steps as illustrated in Figure 3. The following sections provide the details about the adopted methodology:

A. Identify Top Trending Android App

First of all, GPS was explored for available Android game apps. It was observed in the exploration that GPS has a feature called Top Charts (Figure 4) that automatically generates a list of trending Android apps. Therefore, it was found useful to utilize the Top Chart for this study, since it minimizes the extraction processing by directly providing the targeted requirements. Note that the list of Android game apps identified by the Top Chart feature is used in this study. Although, for later steps, a set of queries is designed. However, no additional query is designed to select the trending apps other than GPS's Top Charts feature.

B. Scrape App Details and Reviews

Secondly, after identifying the trending apps, their details and user reviews are scrapped using a free web scraping tool called Parsehub [13]. Although web scraping can be done manually, however, it would become a tedious and challenging process. Therefore, a web crawling tool is used in this study. Among various web scrapping tools, including Scrappy[14], Webhose.io [15], and Content Grabber [16] the simplest tool found was Parse-hub. Parsehub uses web scrapping queries similar to SQL queries, other than that most of its features have drag-and-drop functionality.

To collect the latest data, Parsehub's built-in feature, range selection was utilized, and selected only apps which were published from January 2021 and onwards. After that, a Parsehub query was designed to extract details of the selected apps. The pseudo-code of the query is mentioned by Algorithm 1.

Algorithm 1 The App Details and Reviews Scrapping Pipeline Pseudo-code

```
procedure SCRAPE DETAILS AND REVIEWS($Categories$, $Genres$, $Details$, $Reviews$)
    for all $categories \in TopCharts$ do
        $Categories \leftarrow\ categories$
    end for
    for all $genres \in Categories$ do
        $Genres \leftarrow\ genres$
    end for
    repeat
        for $i \leftarrow 1, n$ do
            $Selectgames, URLs, Ratings, Users, Prices \leftarrow i$
            $Details \leftarrow\ i$
            $i++$ then Goto Reviews
            for $j \leftarrow 1, n$ do
                $SelectDate, Remarks \leftarrow j$
                $Reviews \leftarrow\ j$
            end for
        end for
    until $\neg free_limit$
end procedure
```

The pseudo-code explains the extraction process to first select categories in Top Charts. In each category go through each genre. In each genre, there are lists of Android game apps. Go through each game app and extract their URLs, ratings, users, and prices. After that, go further deep and collect reviews on each app till the free limit of Parse hub allows. The demonstration of the obtained results as extracted categories is provided in Table 1, and genres are shown in Table 2. For more results details, please visit the proposed open-source repository,i.e. TTAG+R.

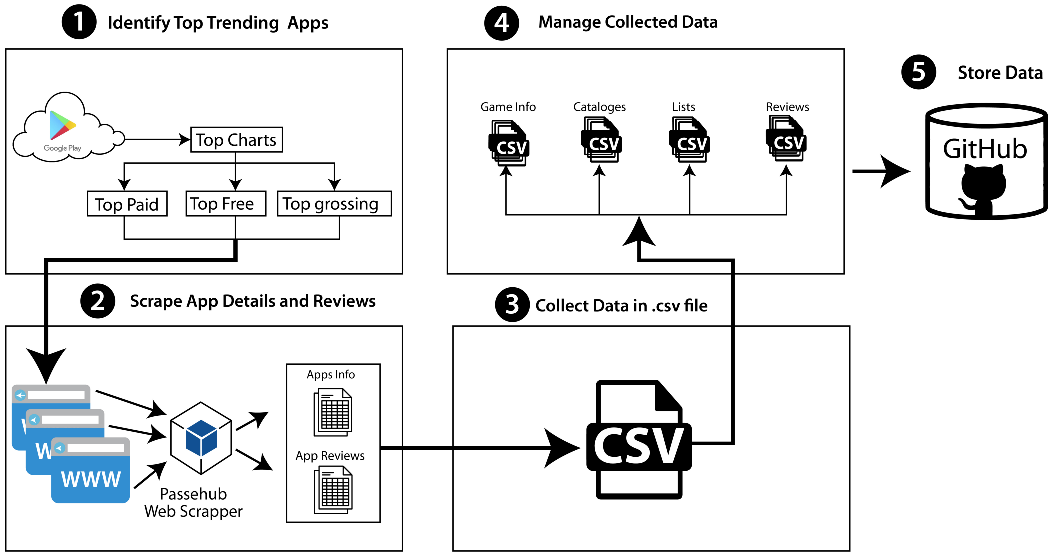
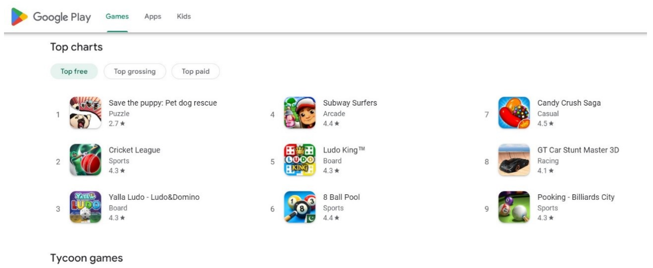


Figure 3. Illustration of Followed Steps to Construct TTAG+R.



Top charts

1 Save the puppy: Pet dog rescue
Puzzle
2.7 ★

2 Cricket League
Sports
4.3 ★

3 Yalla Ludo - Ludo&Domino
Board
4.3 ★

4 Subway Surfers
Arcade
4.4 ★

5 Ludo King™
Board
4.3 ★

6 8 Ball Pool
Sports
4.4 ★

7 Candy Crush Saga
Casual
4.5 ★

8 GT Car Stunt Master 3D
Racing
4.1 ★

9 Pooking - Billiards City
Sports
4.3 ★

Tycoon games

Figure 4. Screenshot of Google Play Store's Top Chart

Table 1. List of Categories of Top Charts

SR	Categories
1	Top Free
2	Top Paid
3	Top Grossing

C. Collect Data in .csv Files

Thirdly, the data was collected and converted into a downloadable .csv file. Importantly, no further filtering of the extracted data was performed. Due to the reason that this dataset is intended to be a complete snapshot of the Google App Store to avoid any restricted filter, which can hamper the general applicability of the dataset.

D. Manage collected Data

Fourthly, a few sub-tables from the collected file were created just to assist users in understanding and reading the information available in TTAG+R. Each table explains the explicit specification of Android apps. The structure and distribution of the tables are shown in Figure 5.

Table 2. List of Genres in Categories

SR	Genres
1	Action
2	Adventure
3	Arcade
4	Board
5	Card
6	Casino
7	Casual
8	Education
9	Music
10	Puzzle
11	Racing
12	Role Playing
13	Simulation
14	Sports
15	Strategy
16	Trivia
17	Word

E. Store Data in Repository

Finally, an open-source repository on GitHub we created. Where stored all the collected and processed data in that repository.

IV. Research Opportunities

This section presents the research opportunities for interested researchers. Following are a few ways in which TTAG+R can assist the research community:

A. Customer Acquisition and Retention

The usability of Android applications is regarded as an emerging area of research. this is mainly because of the

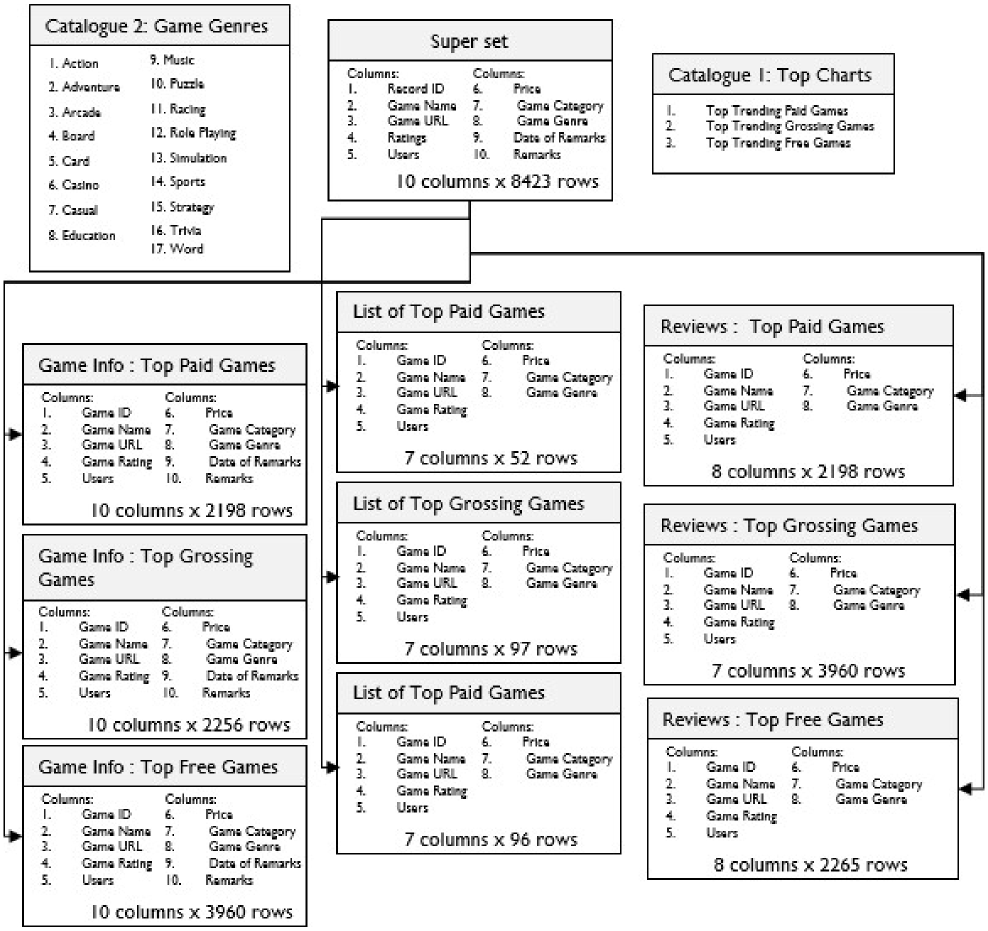

Catalogue 2: Game Genres
1. Action
2. Adventure
3. Arcade
4. Board
5. Card
6. Casino
7. Casual
8. Education
9. Music
10. Puzzle
11. Racing
12. Role Playing
13. Simulation
14. Sports
15. Strategy
16. Trivia
17. Word
Super set
Columns: 1. Record ID 2. Game Name 3. Game URL 4. Ratings 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre 9. Date of Remarks 10. Remarks
10 columns x 8423 rows
Catalogue 1: Top Charts
1. Top Trending Paid Games
2. Top Trending Grossing Games
3. Top Trending Free Games
Game Info : Top Paid Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre 9. Date of Remarks 10. Remarks
10 columns x 2198 rows
Game Info : Top Grossing Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre 9. Date of Remarks 10. Remarks
10 columns x 2256 rows
Game Info : Top Free Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre 9. Date of Remarks 10. Remarks
10 columns x 3960 rows
List of Top Paid Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre
7 columns x 52 rows
List of Top Grossing Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre
7 columns x 97 rows
List of Top Paid Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre
7 columns x 96 rows
Reviews : Top Paid Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre
8 columns x 2198 rows
Reviews : Top Grossing Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre
7 columns x 3960 rows
Reviews : Top Free Games
Columns: 1. Game ID 2. Game Name 3. Game URL 4. Game Rating 5. Users
Columns: 6. Price 7. Game Category 8. Game Genre
8 columns x 2265 rows

Figure 5. Tables and Sub-Tables of the TTAG+R

increasing use of the Android market in the world [17]. The user reviews in TTAG+R can be used to identify usage patterns between different end users. To accomplish this, various types of data analytic techniques can be performed to uncover hidden patterns and relationships. For example, regression analysis is a popular relationship estimation technique [18] that can be applied to TTAG+T to identify trends and patterns. The Cohort is another powerful analytic technique [19] that takes data from a dataset and divides the customers into groups to examine customer behavior concerning the customer life cycle. It can be applied on TTAG+R and valuable insights can be discovered about Android App usage. Likewise, many other techniques can be applied to TTAG+R to attain customer requirements from the dataset, and subsequently, solutions can be proposed for enhancement in the features of the Android app.

B. Attain Advertisements Strategies

The communication and media industries are well-utilizing customers' data and behavior to keep entertaining their audiences [20]. Likewise, the Android industry can use customer remarks to generate content in Android apps attracted to end users. For example, a popular Natural Language Processing (NLP) technique sentiment analysis [21] can be applied to TTAG+R content to generate customer needs. After that, the researchers can recommend improvement guidelines for Android app development.

C. Understanding Market and Competition

TTAG+R can be used to monitor the financial market activities of Android apps. TTAG+R contains three cost-centered categories: (i) Top Free, (ii) Top Paid, and (iii) Top Grossing Android apps. Currently, machine learning [22] and deep learning [23] are gaining the wide attention of researchers to analyze and learn various financial markets. The developer can apply any of the proposed machine learning techniques to TTAG+R to explore and predict the financial aspects in Android Context.

D. Quality App Development

Android app development significantly differs compared to general software development. Hence, traditional quality attributes lack in comprehending the emerging challenges in the context of the Android app. Inspired by this, many researchers have been conducting studies to extract specific attributes from various domains of Android apps. For example, Obaidi et. al [21] extracted app vulnerability attributes. In contrast, Mabrouka et. al [24] reported a list of aesthetic attributes for Android apps. Thus, the researchers can explore TTAG+R for the current state-of-the-art quality attributes for Android Apps. Consequently, the discovered quality attributes support building high-quality Android apps.

Note that the above-mentioned are just a few research directions. However, TTAG+R could be employed in other future research opportunities. Additionally, the developers and researchers could find some other potential areas in the Android apps by exploring the dataset.

V. Limitations and Challenges

In this work, the main encountered challenge was in selecting the source for data extraction. For data extraction, information on Google Play Store's open-access app reviews was selected on the grounds that the Google Play Store has popularity in the Android industry. Moreover, Google Play Store is the official Android publishing platform. Nevertheless, there are other sources available in the market i.e, Amazon or similar archives that may provide more useful insights into Android apps. Therefore, the other sources need to be explored.

Another struggle was faced in finalizing the structure and set of features to be included in the dataset. Due to the reason that there were limited datasets available for Android apps. Thus, it was difficult to follow an appropriate set of guidelines to construct the metadata of the dataset.

One more problem came across in understanding the usage trends for the datasets to depict the clear productivity of the proposed dataset. To achieve usage trends, a thorough literature review was conducted to learn the issues faced in Android games. Once a list of issues was gathered, then concisely each feature was analyzed that can possibly answer the collected list of issues. However, neither the collected issues nor the feature selection was verified by the domain practitioners. Therefore, the analysis has completely relied on the knowledge gathered from the literature study. Finally, similar to the data source, we encountered challenges in selecting the depositing platform for the dataset. Currently, this dataset is deposited in GitHub. The reason to choose the GitHub repository is its popularity among developers. Developers not only deploy their codes in GitHub but also use GitHub as the communication medium in their organization for development purposes. Although, this study has tried to provide all the useful information that can be utilized for Android games-related studies, however, it was discovered later from the practitioner's point of view that the Kaggle is the more in-use repository for ML-based research learning.

VI. Related Work

Software development remains a complex task. In the literature, several researchers have proposed models, frameworks, and strategies for general software development to reduce complexity and risk factors. However, few published studies have focused on presenting models and frameworks for Android app development. For example, there is a dataset for open-source Android Applications developed by Kurtz et al. [25]. The authors represented a code, a developer's perspective, of various open-source Android Applications. However, their presented dataset lacks in considering the user perspectives on Android games. To handle this deficiency, TTAG+R provides data from the user's perspective on the Android Games. There is another dataset for Video Game Development Problems by Politowski et al. [26]. The authors reported quoting various development-related issues while developing a video game by analyzing the postmortems of programmers and experts. It is another dataset solely related to development improvements; however, not for effective planning strategies that can effectively satisfy customers. In 2017, Grano et al. [5] introduced a dataset of Android apps and user feedback. However, the authors have not updated their proposed dataset thereby lacking in accommodating the current knowledge. In contrast, Google Play Store apps is an updated kaggle repository but provide only statistical data on Android apps. However, to the best of our knowledge, there is no reported work that focused on presenting a helpful dataset in providing insight into the context of Android Game development.

VII. Conclusion

This research paper presented a Dataset TTAG+R of Trending Android Games on the Google Play Store. The data is extracted in the form of .csv files. The information lists are provided in raw as well as processed form in the GitHub repository. Briefly, TTAG+R provides a collection of top trending Android games, current state-of-the-art detail descriptions for Android apps, and multiple user reviews of each Android app. Besides, the current study aims to amplify studies that rely on datasets and motivate developers to incorporate users' perspectives in the Android app development life cycle. Moreover, TTAG+R provides a fertile ground for various empirical studies as discussed above in the research opportunity section. Thus, TTAG+R is prepared, organized, and structured in a way that can be quickly processed by analytical queries to assist researchers in shaping the sharp edges of Android apps.

References

[1] D. T. Edi Surya Negara and R. Andryani, "Topic modelling twitter data with latent dirichlet allocation method," pp. 386–390, 2019.

[2] J. Cui, X. Z. Lixin Wang, and H. Zhang, "Towards a predictive analysis of android vulnerability using statistical codes and machine learning for iot applications," *Computer Communications*, vol. 155, pp. 125–131, 2020.

[3] S. H. B. G. N. S. Dehkordi, Mehrdad Razavi and M. H, "Success prediction of android applications in a novel repository using neural networks," *Complex and Intelligent Systems*, vol. 6, p. 573–590, 2020.

[4] A. M. Mouna Hammoudi, Christoph Mayr-Dorn and A. Egyed, "A traceability dataset for open source systems," pp. 555–559, 2021.

[5] F. M. C. A. V. G. C. Giovanni Grano, Andrea Di Sorbo and S. Panichella, "Android apps and user feedback: A dataset for software evolution and quality improvement," pp. 8–11, 2017.

[6] G. Prakash, "Google play store apps," 2021. [Online]. Available: https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps

[7] T. C. Namrud Zakeya, Kpodjedo Ségla and B. B. Alvine, “Probing androvul dataset for studies on android malware classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, pp. 6883–6894, 2022.

[8] Z. D. A. A. X. S. L. L. Xiang Jianwen, Weng Caisheng and T. Jing, “Software aging and rejuvenation in android new models and metrics,” *Software Quality Journal*, vol. 28, p. 85–106, 2019.

[9] A. A. Göksu, İdris and T. Y. Emrah, “Evaluation of mobile games in the context of content: What do children face when playing mobile games?” *SAGE Publications*, vol. 17, p. 388–407, 2020.

[10] “Top trending android games and user reviews.” [Online]. Available: https://github.com/AndroidGamesResearch/Dataset-of-Trending-Android-Games-with-User-Reviews

[11] “Google play store.” [Online]. Available: https://play.google.com/store/games

[12] E. A. A. M. C. Oumayma Hamdi, Ali Ouni and M. W. Mkaouer, “An empirical study on the impact of refactoring on quality metrics in android applications,” pp. 28–39, 2021.

[13] “Parsehub.” [Online]. Available: https://www.parsehub.com

[14] “scrappy.” [Online]. Available: https://scrapy.org/

[15] “Webhose.” [Online]. Available: https://webz.io/data-apis/archived-web-data

[16] “Content grabber.” [Online]. Available: https://contentgrabber.com

[17] S. A. K. Neeraj Mathur and Y. R. Reddy, “Usability evaluation framework for mobile apps using code analysis,” p. 187–192, 2018.

[18] M. K. D. R. B. J. M. Anam Fatima, Ritesh Maurya, “Android malware detection using genetic algorithm-based optimized feature selection and machine learning,” pp. 220–223, 2021.

[19] P. E. J. A. S. K. A. A. D. E. O. S. A. E. M. W. K. Jacob G Scott, Geoffrey Sedor and J. F. Torres-Roca, “Pan-cancer prediction of radiotherapy benefit using genomic-adjusted radiation dose (gard): a cohort-based pooled analysis,” *The Lancet Oncology*, vol. 22, pp. 1221–1229, 2021.

[20] C. V. A. J. S. C. A. V. G. C. Andrea Di Sorbo, Sebastiano Panichella and H. C. Gall, “What would users change in my app? summarizing app reviews for recommending software changes,” p. 499–510, 2016.

[21] G. B. M. D. P. M. L. Bin Lin, Fiorella Zampetti and R. Oliveto, “Sentiment analysis for software engineering: How far can we go?” pp. 94–104, 2018.

[22] M. M. A. Sabeer Saeed and M. Karabatak, “Software engineering for data mining (ml-enabled) software applications,” pp. 1–9, 2021.

[23] D. L. Yanming Yang, Xin Xia and J. Grundy, “A survey on deep learning for software engineering,” *Association for Computing Machinery Journals*, vol. 54, pp. 1–73, 2022.

[24] M. S. Chouchane Mabrouka and K. Ghedira, “The impact of the code smells of the presentation layer on the diffuseness of aesthetic defects of android apps,” *Automated Software Engineering*, vol. 28, pp. 1–29, 2021.

[25] S. A. M. A. R. J. P. A. F. Daniel E. Krutz, Mehdi Mirakhorli and J. Smith, “A dataset of open-source android applications,” p. 522–525, 2015.

[26] G. C. U. J. d. A. W. Cristiano Politowski, Fabio Petrillo and Y.-G. Gu, “Dataset of video game development problems,” p. 553–557, 2020.