Volume: 12 Issue: IV Month of publication: April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61177

Identifying Risky Apps on the Google Play Store: A Decision Tree Approach

Mrs. A. Manga Devi[1] , Beeraka Chaitanya[2], Gunuru Kiran Kishor[3], Talupuri Jahnavi[4], D Durga Bhavani[5] ,P S S Basheeranjali[6]

[1]Assistant Professor , [2, 3, 4, 5, 6]B.tech Students Department of Information Technology, Pragati Engineering College, Surampalem , Andhra Pradesh, India

Abstract: With the rapid advancement of technology, fraudulent phone calls, such as spam and hazardous calls, have become a major issue for the telecom industry, resulting in annual financial losses of millions of dollars worldwide. Phishing and spamming via phone or mobile devices, as well as fraudulent phone calls, have become common dangers to individuals and organizations. Machine learning (ML) and artificial intelligence (AI) have evolved into excellent approaches for detecting and understanding fraudulent or damaging calls. This paper presents an overview of AI-based methods for detecting and evaluating spam or fraud, as well as an evaluation of their shortcomings and potential solutions. A new method for detecting fraudulent calls is proposed, which has shown extraordinary precision and accuracy. A dataset of actual fraudulent calls was used to evaluate the suggested approach. The results also showed that the technique was extremely accurate in detecting possibly fraudulent or spammy calls, as well as malicious calls. Countermeasures can be developed using the knowledge gathered from the research of fraud calls, which showed scammers' methods and approaches.

Keywords: Decision Tree, Google Play Store Apps, Reviews, Ratings, in app purchases, Contains Ad

I. INTRODUCTION

The use of mobile phones has increased as technology advances. The development of different play store apps on various major platforms, like popular Android and iOS, has exploded. It has become a major problem in the business intelligence area due to its rapid daily expansion through its daily use, marketing, and development. The market is becoming more competitive as a result [7]. Companies and software developers are in intense competition with each other to prove their product quality and invest a lot of time and effort in hiring clients to maintain their future success. Customer ratings, and updates on each program they have the opportunity to download play a very important role. This can be a strategy for engineers to identify their flaws and integrate them into the design of a new product that meets the needs of the people. [5].

As a general practice, instead of relying on traditional marketing strategies, under the trees App developers have the ability to majorly support their apps, and ultimately, manage the rank in the App Store. This is often accomplished by using so-called "bot ranch" or "water army" to increase the number of downloads and audits [15].

Occasionally, for the benefit of the developers, they will hire groups of people who will collectively commit fraud and offer spamming comments and ratings on an application. Crowd turfing is the word for this type of behaviour. As a result, it is critical to ensure that consumers are presented with accurate and authentic feedback prior to installing an app in order to minimize errors. In order to overcome and analyse the many comments and ratings that are supplied for

each application, an automated method is required [6]. Because mobile phones are such a common requirement, it is critical that suspicious program be designated as fraud so that play store users may uniquely identify them. It will be impossible for the user to tell if the remarks or ratings that they scroll over are a scam or genuine for their advantage. By offering a comprehensive perspective of ranking fraud detection system, we propose a system that would identify such fake applications on the Play or App store [2]. We can acquire the probability of determining whether an app is fake or not, therefore we present a system that uses four features that are in app purchases, Contains Ad, ratings and reviews to determine the probability of an app whether it's scamming its consumers or not [5]. We start the system by taking into account these four features that are the most crucial in determining the target. Then based on these features the collected data through scraping is being trained through various classification models and then selecting the right and accurate model for the system. In this stage of the selection, we got various models with different accuracy Naive Bayes with 83%, Logistic Regression with 84% and Decision tree with 85% accuracy [2].

II. LITERATURE SURVEY

Nevon Projects has proposed such a comprehensive framework that can be expanded with additional evidence generated by the domain to detect quality fraud. It is one of the most advanced projects to detect fraudulent applications using information algorithms. This program provides only 75-80% accuracy in detecting fraudulent applications.

On paper [7], they provided a complete overview of the facts and a recommended fraud detection framework. They tested three types of verification: Quality-based guarantees, rating based guarantees, review-based validation. In [2] paper, they view only updates as parameters with algorithm naïve bayes. They have developed a system that involves detecting fraudulent applications through emotional commenting and data processing. In [3] paper, first take a look at the app based totally on analytics evaluation as a way to decide whether the software is fake or proper. When enhancing your e-mail, use to test spelling of the file. through the usage of a simple facts set of rules and a manage analyst that exposes app fraud and once more primarily based on feedback from users we provide key-word analysis and based totally on scale. in this manner we also assist with what the consumer thinks approximately our app. on this fraud detection utility, the administrator also offers an app hyperlink that ought to be delivered to the software in which the person when logged in can see the app details and spot the hyperlink for that software. For us in which managers upload the app and offer a hyperlink for that from the Apple keep and Google Play store. to improve our app, we are able to ask the consumer after the usage of our app to get comments, we use this sort of remarks from the consumer and help us enhance the app where the consumer the use of our app will at once see all apps analysed via the administrator and introduced by way of admin. Consequently, the user receives all the records approximately that app and whether the app is fake or not. While an administrator pronounces fraud rather than their fraud, that saves the consumer time and offers person safety.

III. SYSTEM ANALYSIS

A. Existing System

The existing system for the "Fraud App Detection of Google Play Store Apps Using Decision Tree" project involves addressing the increasing need to distinguish between safe and potentially fraudulent mobile applications. With the proliferation of various mobile apps, it becomes essential to assess their safety. The system relies on four key parameters: ratings, reviews, in-app purchases, and the presence of ads within the apps. To make predictions, three machine learning models are employed: Decision Tree classifier, Logistic Regression, and Naïve Bayes. The evaluation of these models is based on four performance metrics: F1 score, Recall, Precision, and Accuracy. A desirable F1 score should exceed 0.7, and a recall score surpassing 0.5 is considered satisfactory, especially when accompanied by higher precision and accuracy. Upon analysis, the Decision Tree model emerged as a robust choice, exhibiting an accuracy of 85%, an F1 score of 0.815, a recall value of 0.85, and a precision of 0.87.

DISADVANTAGES OF THE EXISTING SYSTEM

1) *Limited Feature Set:* The system relies on a relatively narrow set of features, specifically ratings, reviews, in-app purchases, and ad presence. This may not capture all relevant aspects of app behaviour or user interactions, potentially leading to oversights in fraud detection.
2) *Data Imbalance:* The dataset used for training and evaluation might suffer from class imbalance, where the number of fraudulent apps is significantly lower than legitimate ones. This imbalance can affect the model's ability to generalize well to detect fraud accurately.
3) *Static Analysis:* The system appears to focus on static features such as ratings and reviews, but dynamic aspects of app behaviour or changes over time are not considered. Fraudulent apps might evolve, and their characteristics might change dynamically, which could impact the efficacy of the model.
4) *Dependency on User-Generated Content:* Ratings and reviews are user-generated content, and their reliability can vary. Users might provide biased or misleading information, leading to inaccuracies in the model predictions. Additionally, the system might not consider cultural or linguistic nuances that can affect the interpretation of user reviews.
5) *Model Interpretability:* While Decision Tree models are known for their interpretability, they may not capture complex relationships within the data as effectively as more advanced models. The interpretability of the model might limit its ability to discern intricate patterns in the data, potentially affecting the detection of sophisticated fraudulent activities.

B. Proposed System

The proposed system for enhancing the "Fraud App Detection of Google Play Store Apps Using Decision Tree" project aims to address the identified limitations and further improve the accuracy and reliability of fraudulent app detection.

To broaden the feature set, additional dynamic features will be incorporated, considering the evolving nature of mobile applications. This might involve real-time monitoring of app behaviours, tracking changes over time, and analysing user engagement patterns. To mitigate the impact of data imbalance, advanced sampling techniques or ensemble learning methods could be explored. The proposed system will also integrate sentiment analysis and natural language processing to better understand and interpret user reviews, accounting for potential biases and linguistic variations. Moreover, the model architecture will be optimized, possibly exploring more advanced machine learning algorithms or deep learning approaches to capture intricate relationships within the data. Finally, emphasis will be placed on developing a user-friendly interface to facilitate easy interpretation of the model's predictions, promoting transparency and user trust in the fraud detection system. Through these enhancements, the proposed system aims to achieve a more robust and adaptive solution for identifying fraudulent apps on the Google Play Store.

ADVANTAGES OF PROPOSED SYSTEM

Certainly, here are five potential advantages of the proposed system for the "Fraud App Detection of Google Play Store Apps Using Decision Tree":

1) *Enhanced Feature Set:* The incorporation of a more diverse and dynamic set of features, beyond ratings and reviews, enables a comprehensive understanding of app behavior. This enhancement can lead to more accurate fraud detection by capturing a broader range of characteristics associated with both legitimate and fraudulent apps.
2) *Real-time Monitoring:* The proposed system's focus on real-time monitoring allows for the detection of evolving patterns and behaviors in mobile applications. This ensures that the model remains adaptive to changes in fraudulent strategies, providing a proactive approach to fraud detection rather than relying solely on static features.
3) *Improved Model Robustness:* By exploring advanced machine learning algorithms or deep learning approaches, the proposed system aims to enhance the model's ability to capture complex relationships within the data. This can result in a more robust and accurate fraud detection system, especially in scenarios where fraudulent activities exhibit intricate patterns.
4) *Addressing Data Imbalance:* The system's consideration of advanced sampling techniques or ensemble learning methods helps address the challenge of data imbalance. This ensures that the model is trained on a more representative dataset, preventing biases toward the majority class and improving its sensitivity to the minority class of fraudulent apps.
5) User-Friendly Interface: A user-friendly interface contributes to the system's overall effectiveness by making it accessible and interpretable for users. Providing clear insights into the model's predictions enhances user trust and understanding, fostering collaboration between the system and users in identifying potentially fraudulent apps on the Google Play Store.

IV. SYSTEM DESIGN

SYSTEM ARCHITECTURE

Below diagram depicts the whole system architecture.

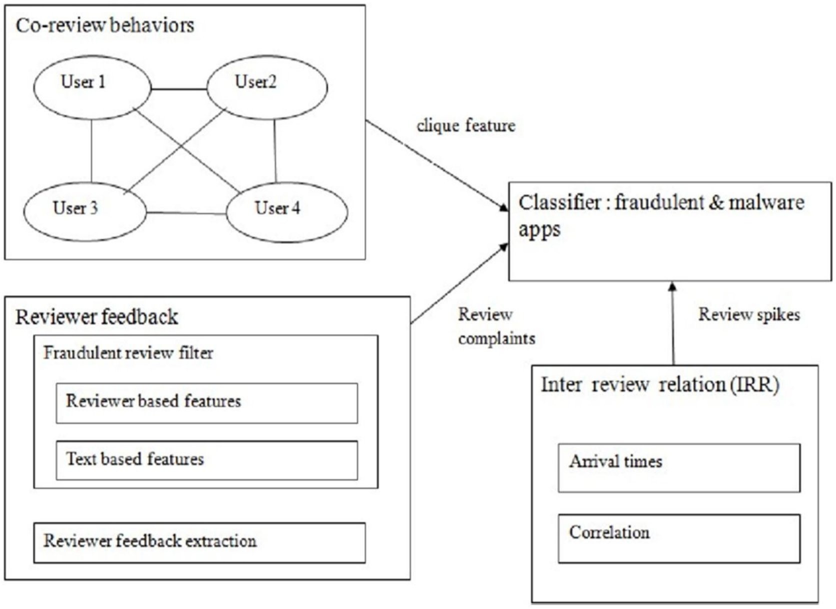


Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

MODULES

1) *Data Collection and Preprocessing:* This module involves the collection of relevant data from the Google Play Store, including app ratings, reviews, in-app purchase information, and ad presence. The collected data undergoes preprocessing to handle missing values, eliminate duplicates, and convert textual information into a format suitable for analysis.
2) *Feature Engineering and Selection:* In this module, the system focuses on enhancing the feature set by considering additional dynamic features that capture the evolving nature of mobile applications. Feature selection techniques may be employed to identify the most relevant attributes for training the fraud detection model, optimizing its performance.
3) *Machine Learning Models:* This module encompasses the implementation and training of machine learning models, including the Decision Tree classifier, Logistic Regression, and Naïve Bayes. Hyperparameter tuning and model evaluation techniques are employed to ensure optimal performance. The Decision Tree model, identified as effective in the analysis, would be a key component.
4) *Real-time Monitoring and Analysis:* To address the dynamic nature of mobile apps, this module involves real-time monitoring of app behaviours, changes, and user interactions. Continuous analysis ensures that the model remains adaptive to emerging patterns and is capable of promptly identifying fraudulent activities as they evolve over time.

User Interface and Reporting: The system includes a user-friendly interface for users to interact with and interpret the model's predictions. This module allows users to input app details or queries and receive clear and interpretable reports on the likelihood of fraud. Visualizations and concise summaries aid in presenting the results, enhancing user understanding and trust in the system.

VI. RESULTS AND DISCUSSION

The sole purpose of the given proposed system is majorly to review the fraud detection of Google play store applications and to use the four parameter methods to differentiate certain fraudulent applications or commonly referred to as spam applications. Experimental analysis is performed on different types of methodology in the proposed manner for the detection of fraud or fake applications. Our system will receive fraud with four types of evidence, such as ad-based ratings, in-app purchases and evidence-based reviews. In addition, the development-based integration approach

incorporates all four aspects to detect fraud. Various machine learning model were implemented which provided different results for the accuracy. By analysis, we found that our given proposed method provides 85% accuracy compared to other algorithms. While independent thinking still exists, the Decision Tree section performs better compared to other models such as the recession and the Naïve Bayes. It is an intuitive algorithm for separation problems. It is a reliable real-time guess, a setback problem. Decision trees can manage non-linear data sets effectively. It plays a role in decision-making in various fields of life, including engineering, social planning, business, and even law

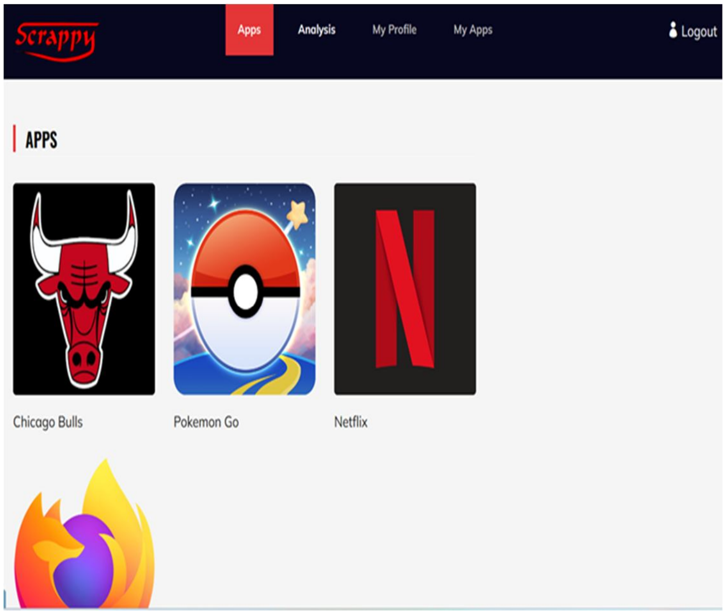
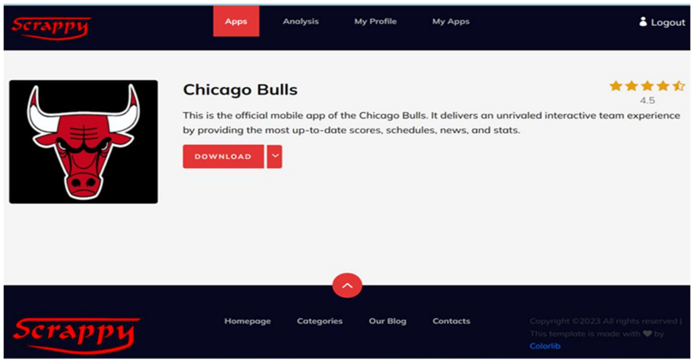


Fig 2. Displaying all available apps in play store



Chicago Bulls

★★★★☆ 4.5

This is the official mobile app of the Chicago Bulls. It delivers an unrivaled interactive team experience by providing the most up-to-date scores, schedules, news, and stats.

DOWNLOAD

Copyright ©2023 All rights reserved | This template is made with ❤ by Colorlib

Fig 3. Downloading Apps from the store based on the rating and reviews

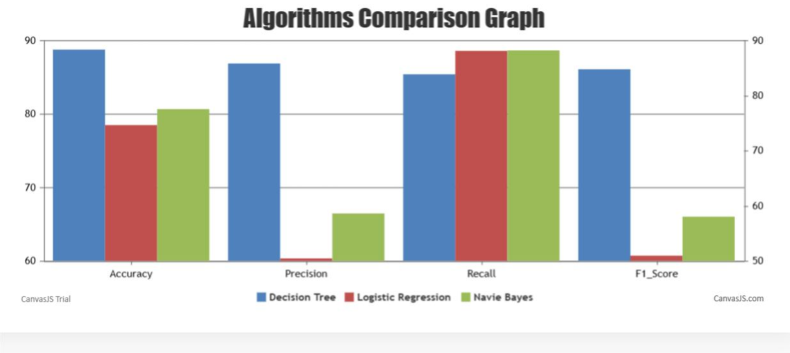
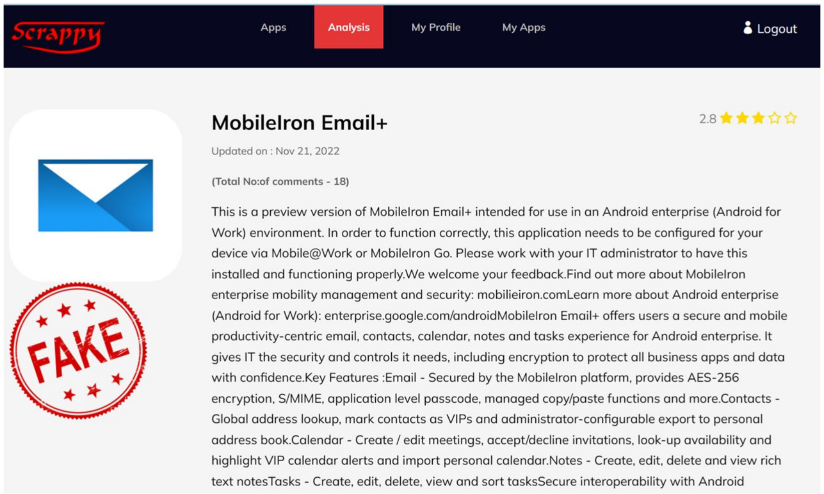


Fig 4. Comparing Different Algorithms Accuracy



MobileIron Email+

2.8 ★★★☆☆

Updated on : Nov 21, 2022

(Total No:of comments - 18)

This is a preview version of MobileIron Email+ intended for use in an Android enterprise (Android for Work) environment. In order to function correctly, this application needs to be configured for your device via Mobile@Work or MobileIron Go. Please work with your IT administrator to have this installed and functioning properly.We welcome your feedback.Find out more about MobileIron enterprise mobility management and security: mobilieiron.comLearn more about Android enterprise (Android for Work): enterprise.google.com/androidMobileIron Email+ offers users a secure and mobile productivity-centric email, contacts, calendar, notes and tasks experience for Android enterprise. It gives IT the security and controls it needs, including encryption to protect all business apps and data with confidence.Key Features :Email - Secured by the MobileIron platform, provides AES-256 encryption, S/MIME, application level passcode, managed copy/paste functions and more.Contacts - Global address lookup, mark contacts as VIPs and administrator-configurable export to personal address book.Calendar - Create / edit meetings, accept/decline invitations, look-up availability and highlight VIP calendar alerts and import personal calendar.Notes - Create, edit, delete and view rich text notesTasks - Create, edit, delete, view and sort tasksSecure interoperability with Android

Fig 5. Providing The data to the application for prediction

VII. CONCLUSION AD FUTURE WORK

In the Era of growing technology, the threat to security is also becoming a major issue and a part of it, today we have a number of apps listed on the Google app stores which also includes various fraud apps that are threat to users' privacy and data. In our model we have worked precisely to detect fraudulent software using 4 parameters including scales, review scores, in-app purchases and content additions. We compared the accuracy of the three algorithms when the resolution tree appeared to be 85% higher. The framework is measurable and can be expanded to further proven domain-based fraudulent evidence. Demonstrated the effectiveness of the proposed system, algorithm detection measurement and standardization of level fraud operations. This can be used effectively to rate fraudulent play store application in play store for security purposes.

REFERENCES

[1] Fawcett, T., Provost, F. Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1, 291–316 (1997).doi:10.1023/A:1009700419189

[2] H. Weng et al., "Online E-Commerce Fraud: A Large-Scale Detection and Analysis," 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 2018, pp. 1435-1440, doi: 10.1109/ICDE.2018.00162.

[3] S. M. Gowri, G. Sharang Ramana, M. Sree Ranjani and T. Tharani,"Detection of Telephony Spam and Scams using Recurrent Neural Network (RNN) Algorithm," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 1284-1288, doi: 10.1109/ICACCS51430.2021.9441982.

[4] Abidogun, Olusola Adeniyi. "Data mining, fraud detection and mobile telecommunications: call pattern analysis with unsupervised neural networks." PhD diss., University of the Western Cape, 2005.

[5] S. Sandhya, N. Karthikeyan, R. Sruthi "Machine learning method for detecting and analysis of fraud phone calls datasets" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

[6] Mohammad Iquebal Akhter, Dr. Mohammad Gulam Ahamad"Detecting Telecommunication fraud using neural networks through data mining" international Journal of Scientific & Engineering Research, Volume 3, Issue 3, March-2012.

[7] I. Murynets, M. Zabarankin, R. P. Jover and Panagia, "Analysis and detection of SIMbox fraud in mobility networks," IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, Toronto, ON, Canada, 2014, pp. 1519-1526, doi: 10.1109/INFOCOM.2014.6848087.

[8] Crawford, M., Khoshgoftaar, T.M., Prusa,J.D. et al. Survey of review spam detection using machine learning techniques. Journal of Big Data 2, 23 (2015).doi:10.1186/s40537-015-0029-9.

[9] Marzuoli A, Kingravi H, Dewey D and Pienta R. (2016). Uncovering the Landscape of Fraud and Spam in the Telephony Channel 2016 15th IEEE International Conference on Machine Learning and Applications(ICMLA). 10.1109/ICMLA.2016.0 153. 978-1-5090-6167-9. (853-858).

[10] B. Teh, M. B. Islam, N. Kumar, M. K. Islam and U. Eaganathan,"Statistical and Spending Behavior based Fraud Detection of Card-based Payment System,"2018 International Conference on Electrical Engineering and Informatics (ICELTICs), Banda Aceh, Indonesia, 2018, pp. 78-83, doi: 10.1109/ICELTICS.2018.8548878.

[11] H. Tu, A. Doupe, Z. Zhao, and G.-J. Ahn, " Sok: Everyone hates'robocalls: A survey of techniques against telephone spam," 2016 IEEE Symposium on Security and Privacy (SP), pp. 320 338, 2016.

[12] M. Crawford, T.M. Khoshgoftaar, J.D Prusa, A.N. Richter, H. Al Najada, "Survey of review spam detection using machine learning techniques", Journal Of Big Data, 2, pp. 1-24, 42015.