Fraud App Detection Using Sentiment Analysis

Under the Guidance of
Prof. Suvarna. S. Wakchaure,
Department of Computer Engineering
Sir Visvesvaraya Institute
Of Technology, Nashik

Miss. Kalyani Babaji Suryawanshi,
kalyanisuryawanshi39@gmail.com,
Department of Computer Engineering
Sir Visvesvaraya Institute
of Technology, Nashik

Miss. Shruti Sanjay Tungar,
tungarshruti04@gmail.com,
Department of Computer Engineering
Sir Visvesvaraya Institute
Of Technology, Nashik

Miss. Shruti Rajendra Jadhav,
shrutijadhav132004@gmail.com.,
Department of Computer Engineering
Sir Visvesvaraya Institute
of Technology, Nashik

Miss. Aditi Bharat Vishwas
aditivishwas614@gmail.com,
Department of Computer
Engineering
Sir Visvesvaraya Institute
Of Technology, Nashik

---***---

Abstract - The smartphone industry is expanding quickly, there are increasing numbers of mobile applications available in the market every single day.Users have confusion while downloading applications because there are a lot of them on the market.As a result, it's important to monitor and create a system to determine whether or not the apps that are there are authentic.The goal is to create a system that uses sentiment analysis and the Naive Bayes algorithm to identify fraudulent apps before users download them.Therefore, we are putting forth a system that uses the Naive Bayes algorithm to evaluate the application's data, comments, and three reviews in order to produce results.Consequently, choosing a fraud application will be simple.

Keywords: - Naive Bayes, Sentimental Analysis, Review based proof, positive negative ratings, Rate evidence, Users satisfaction, Suspicious.

1.INTRODUCTION

The mobile application has become widely accepted and popular as a result of the quick development of mobile technology. Ranking fraud is the main issue facing the mobile app industry because of the sheer volume of available applications.The mobile users choose focusing their attention on the best bother rank applications that are willing to bring them with superior quality and satisfaction. Within the huge app market, high-rated apps play a significant function in the user's choice on looking to their preference for positive reviewed experiences, whereby individuals go for was well rated below. The success of the top-ranked apps and the popularity amongst users not only justifies the application of user feedback to influence the mobile preferences, but it also proves the value that the likes of Foursquare or Hollywood Vampire have rendered us. Smartphone users need to visit app stores, including Apple's, Google Play To Store, and some others, to download the relevant apps.Every user may view the different application listings when he visits the Play Store.The foundation of this list is marketing and advertising.The user often downloads applications from a list

without knowing which ones are useful or useless, leading to fraud in mobile application lists.To prevent this, an application is created to list these applications.The project investigates review-based evidences using historical records and uses a Naive Bayes algorithm for fraud detection.The evaluation of the system incorporates the use of real-world app data from the Google App Store.

II RELATED WORK AND LITERATURE SURVEY.

This project's primary focus is on using data mining and sentiment analysis to extract the generated dataset. Through this process, we can establish the actual value of the apps that are accessible on both the Play and App stores. A Set up of this kind would need handling enormous amounts of data, and data mining in addition to visual data would be helpful in implementing the system. Business oriented information mining is the process of obtaining critical data from large databases. This procedure involves structuring the information at a later stage, which, in its turn, defines what information mining is. This process includes sentiment analysis as a component. Since it is the method for looking at explanations and deriving abstract information from them. It's identifying the extremes of the announcements at a very fundamental level. Data is collected via many online sources, mobile apps, and exchanges that include questionnaires, comments, and other information specific to each company. In addition, sentiment analysis is used to deconstruct the data in order to prepare upgrades in the future based on measurements obtained from estimation research. Examining large-scale informational collections is an important yet problematic task. Data representation techniques might be useful in resolving the problem. Investigating visual

information has a lot of possibilities and uses. One such use is the detection of misrepresentations. Additionally, data representation innovation will be used by information mining to provide a better evaluation of the information [1]. In this research, we suggest and apply the effective use of data mining for fraud determination. We may decide our backend data retrieval more easily by using different data mining approaches and algorithms. There are several categories into which fraud falls. [2] These are the applications of data mining. To facilitate grouping, extortion has been divided into four main categories: safety, media communications blackmail, computer sabotage and financial manipulation. deceptive display. Related extortion, which includes banking deception, securities and wares extortion, and other forms of related extortion such as fiscal report extortion, citizen extortion, and word-related fraud, constitute a further category under financial extortion, while medical fraud, crop protection extortion, and accident protection extortion constitute another group under insurance fraud. One of the previous literature surveys that was carried over also used the mobile user's IP address. [3] The phrase "misrepresentation application" is becoming more and more common in portable application advertisements. These days, in the portable market, anticipation and recognition play a critical role. The application of The Fraud Ranking System is recommended for the case of extortion audit in the single client system, which is flexible. Applications are allocated a spot based on their survey evaluation accumulation. While it is able to identify the unique aspects of the sources, in terms of effectiveness, it falls short owing to the fact that IP spying is possible. Users are able to rate an app many times and modify their IP address because to this IP spying. The star ratings that are offered for each and every application are insufficient to decide whether or not the app is appropriate to download on a mobile device. It is not entirely correct to believe in star ratings, as stated in [16], as developers themselves have the ability to modify them. Reading reviews is valued more highly than ratings. In general, it is suggested [17] to look via more reputable sources, including carefully curated reviews from third parties or looking through the developer's previous applications. gathering a particular app dataset over time and classifying the reviews as either good or negative [ev]. The efficiency of using such a model is also reduced even in the popular feedback (the N-gram model with N=2), which, as stated earlier, you can use faster using fewer words. The proposed approach simplifies the classification of words regarding their concreteness.

III EXPERIMENTAL SYSTEM

Failures:

There are a lot of fraud app detection technologies available these days that use rating evidence to identify phony applications in a variety of ways. This module computes the average evaluation of a certain application and compare it to a threshold. Generally, ratings range from one to five. Ratings that fall between two and three are regarded as negative ratings, while those that rise over three are regarded as positive ratings. Ultimately, the output takes the form of ones and zeros; that is, an output of zero is generated with a negative rating, and an output of one with a good rating. According to the proposed system, There are instances where the user reviews and the numerical rating diverge greatly. The Naive Bayes technique is used to refute this claim. Sentiment analysis aids in identifying the emotional undertones of words used in online communication. This technique aids in searching for emotive remarks left by users on certain programs. The computer may learn from a training data set and analyser attitudes and emotions regarding reviews and other texts by using sentiment analysis. The right Android application may be chosen and whether or not there is fraud by using emotional analysis to reviews and comments.
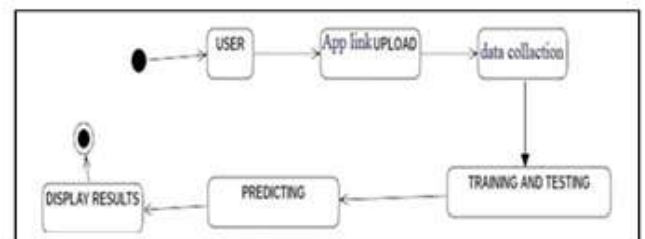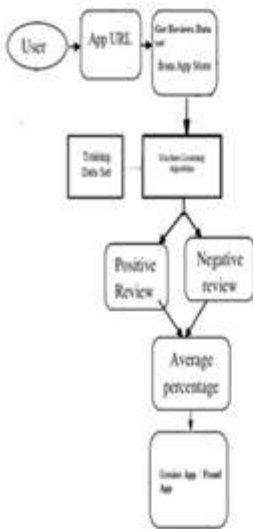

Fig. 1 System Architecture Diagram

Fig. 2 Admin Activity Diagram

IV IMPLEMENTATION

Developing a code that is simple to read and comprehend is the aim of implementation. This is the most important phase in creating programming that works or a framework that convinces the customer that the new programming works and produces results that are worth their while. For testing, debugging, and modification to be simple, the source code has to be understandable. since they make up a sizable amount of software expenses. Code quality, error correction, and performance are all addressed in exact implementation. Coding style approaches, standards, and recommendations are all part of this phase.

Data Collection (Crawler): Using the Google Play Store app and retrieving data based on the provided app ID. App name and review from the user

Grouping Reviews: These are the steps that make up this phase:

Tokenization: This involves segmenting a text stream into words, phrases, symbols, or meaningful Components referred to as tokens. The token list is used as the input for additional processing.

Eliminating stop words involves taking out often used words like "a," "the," "and," "for," "from," "is," "in," and many more.

Stemming: To locate the base word, utilise the stemming method. The basic word algorithm is the Porter Stemmer Algorithm.

Sentiment Analysis: Determine the sentiments of the reviews following the system's preprocessing. It will provide a
favourable or negative rating to the review. The review's attitude, whether favourable or negative, will be determined by the system.

Aggregation: Following Sentiment Analysis pre-processing a favourable review raises the score by one; a bad review raises the score by one.

This will allow it to calculate the average score for each review, identify good and negative reviews, display them on a graph, and evaluate if the app is fraudulent or not depending on the percentage of positive and negative reviews. No.

Algorithm:

In machine learning, there is a family of simple probabilistic classifiers known as "naive Bayes classifiers" which are based on using the Bayes theorem under the strong (naive) assumption of feature independence. There has been deep research on Naive Bayes since the 1950s. It was presented. Albeit under a different name, into the text retrieval community in the early 1960s[1]: Notably, 488 is still a prevalent (benchmark) method of text categorization, the problem of classifying documents into one of several categories (spam or legitimate, sports or politics, etc.), based only on word frequencies as the features. The support vector machines and other advanced techniques are competitive in this domain if they are well pre-processed. It is also used in automatic medical diagnosis. Naive Bayes classifiers, as they require a linear number of parameters in the (features/ predictors) number of variables in a learning task, are highly scalable.

As with many other types of classifiers, most-probability preparation of was done in a linear time by checking the closed form expression instead of iteratively approximating. Naive Bayes models are referred to by several names in the literature on statistics and computer science, such as independence Bayes and simple Bayes. While the Bayes theorem is used in the decision process of the classifier, it is not (obviously) a Bayesian method. A simple way to construct classifiers is called a naive Bayes.In these models, issue cases are encoded as vectors of feature values and class labels are assigned based on a threshold on class labels. Instead, these classifiers are trained on a set of algorithms. The general idea is that for every class variable, the value of each feature is considered to be independent from the value of any other feature. For instance, if a red, round, and roundish fruit of diameter about 10 cm is an apple. If there is a correlation between the colour, diameter, and diameter, the classifier regards them as having separate effects on the likelihood that this fruit is an apple. In supervised learning environments, Naive Bayes classifiers can be taught very well for some types of probability models, and the maximum likelihood approach is applied in the real- world parameter estimation in the Naive Bayes models. That is, you can

employ the Naive Bayes model without applying any Bayesian techniques, and it has performed well in a lot of complicated real-world cases. Naive Bayes classifiers may seem incredibly efficient, but there are good theoretical reasons for this. According to a 2004 study on the Bayesian classification problem, a rigorous comparison of various classification algorithms performed in 2006 showed

V RESULTS AND DISCUSSIONS

In this work, the idea of sentiment analysis was used to identify fraud applications. The architectural diagram, which provided information on the algorithm and procedures used in the project, supported it. After data is gathered, it is assessed using the established supporting algorithms. This is a novel method where the evidence is condensed and combined into a single outcome. The suggested system for rating fraud detection is scalable and adaptable to other domain-generated information. In this project, a mobile app fraud detection system has been built. To be more precise, it first retrieves the information about the chosen app from the Google Play Store's history records.

Then it determined evidence based on reviews to detect fraud. Additionally, it suggested using the Naive Bayes technique to assess the veracity of leading sessions from mobile applications. Last but not the least, it performs extensive tests with the dataset containing the real App data from Google Application Store to validate the solution proposed. Then, performance of the proposed method was inferred through experiment. From the results of the testing it was established that it is possible to scale up the detection algorithm while the proposed approach proved to be effective and some regularity was found in the ranking of fraudulent actions. Future research is intended to examine more potent fraud evidence and examine the hidden connections between ratings, reviews, and rankings. In addition, for the purpose of improving the user experience, it will be expanded to include a rating fraud detection technique along with new services related to mobile applications that include mobile app recommendations.

Advantages:
   1. Extremely fast feature computation.
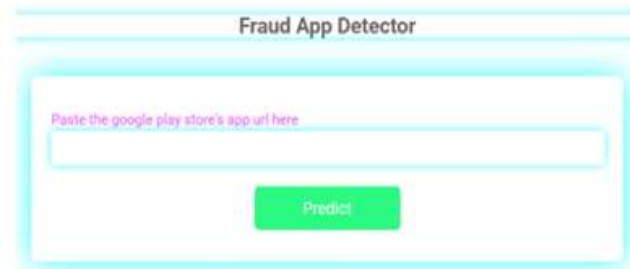   2. Efficient feature selection.

Fig. 3.1 HOME PAGE

Fig. 3.2 SAFE APP DETECTION

Fig. 3.3 MIGHT NOT BE VERY SAFE

Fig. 3.4 FRAUD APP DETECTION

VI CONCLUSION

In this project, a system for detecting fraud in mobile application ratings has been developed. In particular, it first demonstrated that leading sessions were the site of ranking fraud and offered a way to mine leading sessions from an App's past ranking data. Then, in order to detect ranking fraud, it recognised evidence based on ranking, evidence based on rating, and evidence based on review. This technique has a distinct advantage in that all the information can be represented using sentiment analysis, making it simple to expand with additional evidence derived from domain expertise in order to identify ranking fraud. The efficacy of the proposed strategy was demonstrated by experimental data. The detection of fraud apps was achieved through the identification of evidence through a review process

Future Scope:

 Future work should attempt to combine our technique with other modalities such as audio modality, including working with other data sets.

REFERENCES

1) Esther Nowroji., Vanitha., "Detection Of Fraud Ranking For Mobile App Using IP Address Recognition Technique", vol. 4, International Journal for Research in Applied Science & Engineering Technology, 2016.

2) FuzailMisarwala, KausarMukadam, and KiranBhowmick, "Applications of Data Mining in Fraud Detection", vol. 3,2015.

3) Daniel A. Keim, "Information Visualizing and Visual Data Mining" IEEE Trans. Visualization and Visual Data Mining, vol. 8,Jan-Mar 2002. (references)

4) Ahmad FIRDAUS, Nor Badrul ANUAR, Ahmad KARIM, MohdFaizalAb RAZAK, "Discovering optimal features using static analysis and a genetic search based method for Android malware detection" Frontiers of Information Techonology and Electronic Engineering, 2018.

5) JavvajiVenkataramaiah, BommavarapuSushen, Mano. R, Dr. GladispushpaRathi, "An enhanced mining leading session algorithm for fraud app detection in mobile applications" International Journal of Scientific Research in Engineering., April2017.

6) Avayaprathambiha.P, Bharathi.M, Sathiyavani.B, Jayaraj.S "To Detect Fraud Ranking For Mobile Apps Using SVM Classification" International Journal on Recent and Innovation Trends in and Communication, February2018 Computing vol. 6,

7) Suleiman Y. Yerima, SakirSezer, Igor Muttik, "Android Malware Detection Using Parallel Machine Learning Classifiers", 8th International Conference on Next Generation Mobile Applications, Technologies,Sept.2014. Services and

8) SidharthGrover,"Malware detection: developing a system engineered fair play for enhancing the efficacy of stemming search rank fraud", International Journal of Technical Innovation in Modern Engineering October2018 &Science,Vol. 4,

9) PatilRohini, Kale Pallavi, JathadePournima, KudaleKucheta, Prof.PankajAgarkar,"MobSafe: Forensic Analysis ForAndroid Applications And Detection Of Fraud Apps Using CloudStack And Data Mining", International Journal of Advanced Research in Computer Engineering &Technology,Vol. 4, October2015

10) Neha M. Puram, Kavita R. Singh,"Semantic Analysis of App Review for Fraud Detection using Fuzzy Logic", International Journal of Computer & Mathematical Sciences,Vol. 7, January2018

11) VivekPingale, LaxmanKuhile, Pratik Phapale, Pratik Sapkal, Prof. Swati Jaiswal,"Fraud Detection & Prevention of Mobile Apps using Optimal Aggregation Method", International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 8, March2016.

12) D.Janet, Vikrant Chole ,"A Review on Ranking Based Fraud Detection in Android Market", International Journal of Science and Research,Vol. 6, January2017.

13) MahmudurRahman, BogdanCarbunar, MizanurRahman, and DuenHorngChau,"Search Rank Fraud and Malware Detection in Google Play", IEEE Transactions on Knowledge and Data Engineering,Vol. 29, June2017.

14) Tahura Shaikh#1, Dr. DeepaDeshpande,"Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews", International Journal of Computer Trends and Technology (IJCTT),Vol. 36, June2016. (Online) Available:https://www.makeuseof.com/tag/who invented-the-first computer/.(Online) Available: https://lifehacker.com/why-you shouldnttrust- app-store-reviews-and-what-to 1515379780

15) Vinothan, D.,Saravanan, M.," Institution System analysis by using similarity based clustering on social network access ",Pakistan Journal of Biotechnology. Vol 13,2016,pp 1-4.

16) Monika Pandey, Prof. TriptiSharma,"Fraud App Detection using Fuzzy Logic Model Based onSentiment of Reviews", International Research Journal of Engineering and Technology,Vol. 5, Sep2018.

17) Gladence, L. Mary, M. Karthi, and V. Maria Anu. "A statistical comparison of logistic regression and different Bayes classification methods for machine learning." ARPN Journal of Engineering and Applied Sciences 10, no. 14 (2015): 5947-5953.

18) Dr.R.Subhashini and Akila G,"Valence arousal similarity based recommendation services ", IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2015.