Detecting Fraud Apps using Sentiment Research

Mandava Rama Rao, Nandhini Kannan, CH V S Nihanth

***ABSTRACT---** With the increase in the number of mobile applications in the day to day life, it is important to keep track as to which ones are safe and which ones aren't. One can't judge how safe and true each application is based only on the reviews that are mentioned for each application. Hence it is a need to keep track and develop a system to make sure the apps present are genuine or not. The objective is to develop a system in detecting fraud apps before the user downloads by using sentimental analysis and data mining. Sentimental analysis is to help in determining the emotional tones behind words which are expressed in online. This method is useful in monitoring social media and helps to get a brief idea of the public's opinion on certain issues. The user cannot always get correct or true reviews about the product on the internet. We can check for user's sentimental comments on multiple application. The reviews may be fake or genuine. Analyzing the rating and reviews together involving both user and admins comments, we can determine whether the app is genuine or not. Using sentimental analysis and data mining, the machine is able to learn and analyze the sentiments, emotions about reviews and other texts. The manipulation of review is one of the key aspects of App ranking fraud. By using sentimental analysis and data mining, analyzing reviews and comments can help to determine the correct application for both Android and iOSplatforms.*

***Keywords** — Sentimental Analysis, Data mining, Review based evidence, positive negative ratings, Rate evidence, Users review, Leading session*

I. INTRODUCTION

With the growth in technology, there is an increase in the usage of mobiles. There has been a vast growth in the development of various mobile applications on numerous platforms such as the popular Android and iOS. Due to its rapid growth day by day for its everyday usage, sales and developments, it has become a significant challenge in the world of the business intelligence market. This gives rise in the market competition. The companies and application developers are having a tough competition with one another in order to prove their quality of product and spend an immense amount of work into attracting customers tosustain their futureprogress.

The most important role that plays is the customers ranking, ratings and reviews on that specific application which they happen to download. This could be a way for the developers to find their weakness and enhance into the development of a new one keeping in mind the peoplesneed. Notonlythat,certaintimesguiledevelopersmisleadinglythe recognition of their apps or malicious ones use it as a platform to spread malwarethroughout.

As an ongoing pattern, rather than depending on customary promoting arrangements, under the trees App developer's option in contrast to some false way to intentionally support their Apps and in the long run controls the outline rankings on an App store. This is generally executed by utilizing so-called "bot ranches" or "human water armed forces" to expand the Application downloads evaluations and audits in an exceptionally brief time.

Certain times, just for the upliftment of the developers, they tend to hire teams of workers who commit to fraud collectively and provide false comments and ratings over an application. This is known to be termed as crowd turfing. Hence it is always important to ensure that before installing an app, the users are provided with proper and genuine comments in order to avoid certain mishaps. For this, an automated solution is required to overcome and systematically analyse the various comments and ratingsthat are provided for eachapplication[19,20].

With mobile phones being a quite popular need, it is essential that suspicious applications must be marked as fraud in order to be identified by the store users. It will be difficult for the user to determine the comments that they scroll past or the ratings they see is a scam or a genuine one for their benefit. Thereby, we are proposing a system which will identify such fraudulent applications on Play or App store by providing a holistic view of ranking fraud detection system.

By considering data mining and sentiment analysis, we cangetahigherprobabilityofgettingrealreviewsandhence we propose a system that intakes reviews from registered users for a single product or multiple and evaluate them asa positive or negative rating [18]. This can also be useful to determinethefraudapplicationandensuremobilesecurityas well.

We initiate the system by considering the mining leading session or also the active periods of the applications. This influences in detecting local anomaly than the global anomaly of the app ranking. In particular, in this, we first propose a basic yet fruitful calculation to recognize the leading sessions of each App dependent on its authentic positioning records. At this point, of the investigation of Apps' positioning practices, it finds the fake Apps that regularly have distinctive positioning examples in each driving session contrasted and ordinary Apps.

Furthermore, we inspect through three types ofevidences namely ranking based, rating based, and

Revised Manuscript Received on July 10, 2019.

Mandava Rama Rao, Department of Computer Science & Engineering SRM Institute of Science and Technology, Chennai, T.N, India. (E-mail: ramarao14981@gmail.com)

Nandhini Kannan, Department of Computer Science & Engineering, SRM Institute of Science and Technology, Chennai, T.N, India. (E-mail: knandhini98@gmail.com)

CH V S Nihanth, Department of Computer Science & Engineering SRM Institute of Science and Technology, Chennai, T.N, India. (E-mail: chnihanth98@gmail.com)

Retrieval Number: B11070782S319/19©BEIESP
DOI : 10.35940/ijrte.B1107.0782S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

review based by modelling the consolidation of the three through statistical hypothesestests.Regardless,thepositioning-basedevidences can be influenced by the App developer’s status and some genuine advertising efforts such as the “constrained-time markdown”. Thus, it is inadequate to consider just the rank-basedconfirmations.

Along with this, the proposed system introduces two sorts of extortion evidences dependent on Apps rating and survey history which mirror some unique patterns from Apps. Also, an aggregated method is utilized to integrate all the evidences that are necessary for detecting fraud. In order to do so, we evaluate the proposed system by using real- world application data collected from google play store and iOS app store for a long period of time.

The paper is divided into V sections which are organized as Section II which describes Literature Survey; Section III mentions System Architecture and in Section IV discusses the working, framework and algorithms used. Finally, Section V briefs about the conclusion.

II. LITERATURESURVEY

The main focus of this project is upon the sentiment analysis and data mining to extract the dataset produced. By using this method, we will be able to determine the true valueoftheapplicationswhichareprovidedinPlayandApp stores. Such a proposed system will contain a huge amount of data set that has to be dealt with and using data mining alongwithvisualdatawillhelpincarryingoutthesystem.

Information or data mining is the way toward extricating required information from substantial informational collections and changes it into a justifiable arrangementforsometimelater,essentiallyutilizedforsome, business based reason. Sentiment Analysis is pitched into this procedure as a piece of it. Since it is the way toward examining explanations and acquiring abstract datafrom them. At an exceptionally fundamental dimension, it is discovering extremity of the announcements.

Information is gathered from different internet-basedlife, portable applications and exchanges which contain surveys, remarks and different data identified with the individual business. Further here feeling examination is utilized for breaking down the information for future upgrades dependent on the measurements acquired by estimation investigation.

The investigation of extensive informational collections is a critical however troublesome issue. Data representation procedures may help to take care of the issue. Visual information investigation has high potential and numerous applications, for example, misrepresentation discovery also, information mining will utilize data representation innovation for an improved information examination [1].

Data mining is utilized in determining fraud efficiently and that’s what we propose and implement in this paper. By utilizing various data mining techniques and algorithms, it would become easier for us to determine our backend retrieval of data. Fraud can be classified into various types [2] which are the applications of data mining. With the end goal of grouping, extortion has been separated into four general classifications budgetary misrepresentation, media communications extortion, PC interruption and protection misrepresentation. Budgetary extortion is additionally separated into bank misrepresentation, securities and wares extortion and different kinds of related extortion which incorporates fiscal report extortion, citizen extortion and word related misrepresentation, while Insurance extortion is additionally ordered into medical coverage misrepresentation, crop protection extortion and accident protection extortion.

Using the IP address of the mobile user were also one of the earlier literature surveys which was carry forwarded. [3] In the portable application advertise, the term called misrepresentation application is getting prevalent. In nowadays, recognition and anticipation are assuming a crucial job in the portable market. For the identification of extortion audit to the single client framework (i.e.,versatile), the Fraud Ranking System is proposed. Evaluations are accumulated to give a position to each application. Although it had identified the sources uniqueness it wasn’t quite efficient considering the fact that IP snooping can be done. This IP snooping allows the users to change their IP address and allow them to rate an app more thanonce.

The star ratings which are provided for every single application isn’t quite enough in determining whether the app is suitable to be loaded on the mobile or not. As described [16] that it’s not quite right to believe into star ratings as they can be manipulated by the developers themselves. It is considered into reading the reviews more than ratings. Generally, it is advised [17] to check more reliable sources such as curated third part reviews or checking the developer’s other apps.

Collection of a specific app dataset for a period of time and differentiating them as positive and negative reviews [ev]. Utilizing fewer words in the reviews, that is, using the N-gram model (N=2) is more efficient for the accuracy of semantic classification. Lesser the words, it is easier to classify them according to their category as the proposed system.

III. SYSTEMDESIGN

From the Literature survey and other past proposed systems which were developed for this very purpose, the problem in eradicating the fraud application is still under work. There are certain works that involve the usage of web ranking spam detection, online review spam and mobile applicationrecommendationorevenfocusesonthedetection of malwares in the apps before downloading them. Google uses FairPlay system which is able to detect the malwares that are present in certain apps only but haven’t been efficient enough to do so due to the concealing properties. The user can be tricked into downloading an application by its ratings even when it does contain certain viruses that can affect the functioning of themobile.

Although there has been other existing systems, the main focus isn’t just on recommendation or spam removal. Some of the approaches can be used for anomaly detection from the historical rating and review records but they aren’t efficient enough to extract fraud evidences for a certain time period (mining leading sessions). Which the wide growth of apps in stores, it becomes a cumbersome task to determine whichofthemaregenuineornotbasedontherankingalone.

Here we propose a system which involves in detecting the fraud apps using sentient comments and data mining.We areabletochecktheuser'ssentimentalcommentonmultiple applicationsbycomparingthereviewsofadminandtheuser. By looking into these comments, we are able to distinguish them as positive or negative comments. With the aggregationsofthreeevidences:rankbased,ratingbasedand review based we are able to get a higher probability ofresult. The data is extracted and processed by the mining leading sessions. The data is then evaluated on the three mentioned evidences and are concatenated before the end result. It is vitaltobriefaboutsentimentanalysisanddataminingbefore continuingfurtherintotheproposedsystemandalgorithms.

A. SentimentAnalysis

Sentiment Analysis also known as Opinion mining is a relevant mining of content which recognizes and extricates emotional data in the source material and helping a business to comprehend the social slant of their image, item or administration while observing on the web discussions. Sentiment analysis is the most widely recognized content grouping device that investigations an approaching message and tells whether the basic estimation is sure, negative or unbiased.

At present, sentiment analysis is a theme of incredible intrigue and advancement since it has numerous handy applications. Since freely and secretly accessible data over theInternetiscontinuallygrowing,countlesscommunicating conclusions are accessible in audit locales, discussions, online journals, and web-based socialnetworking.

With the assistance of opinion mining frameworks, this unstructured data could be consequently changed into organized information of popular assessments about items, administrations, brands, governmental issues, or any point that individuals can express feelings about. This information can be exceptionally valuable for business applications like showcasing examination, advertising, item surveys, net advertiser scoring, item criticism, and client administration.

There are numerous sorts and kinds of opinion mining and tools run from frameworks that attention on the extremity (positive, negative, unbiased) to frameworks that recognize sentiments and feelings (irate, glad, miserable,and so forth) or distinguish aims (for example intrigued v. not intrigued).

B. Data Mining

There is an immense measure of information accessible in the Information Industry. This information is of no utilization until it is changed over into helpful data. It is important to examine this gigantic measure of information and concentrate helpful data from it. Extraction of data isn't the main procedure we have to perform; information mining additionally includes different procedures, for example, Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

When every one of these procedures is finished, we would most likely utilize this data in numerous applications, for example, Fraud Detection, Market Analysis, Production Control, Science Exploration, and so on. Data mining is utilized here to look into the review data by the apps. This data is then filtered and processed before it can go through the process of sentiment analysis. The reviews are extracted and distinguished based on various datasets that are in the database.Accordingly,thetextisevaluated.Tobeparticular, we are using text data mining which is also referred as text mining. From the texts which are extracted(reviews) it is easiertoanalyzewordsoraclusterofwordsthatareused.

C. Architecture Diagram

Our proposed system as in Fig 1 gives an overall flow of the process which is happening. It begins with the extraction of data that is the historical records of the applications and userdetailsfromthestore.Theadminaddsanewapplication to the database along with the rating details. From here, it will mine the leading session where it is calculated on the basis of evidences observed for that particular app. For this, the mining leading session algorithm is used which is able to identify the leading session andevents.

After that, the evidences of rating, ranking and reviews are looked into one by one. The estimation of these evidences would be assembled with the thought of the different time sessions, essentially dependent on the main sessions. Positioning based confirmations are the one which is finished by the application head board to give a superior survey of applications to the clients utilizing cell phones.

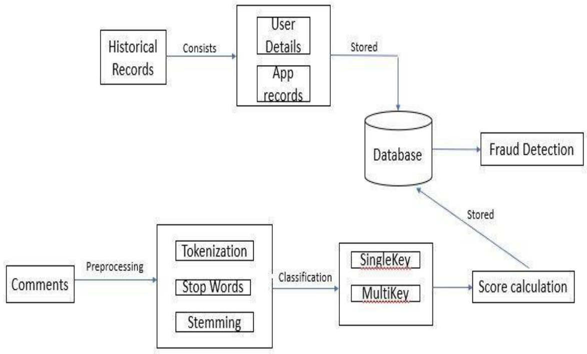

Fig 1. System Architecture

The ranking of applications would comprise of three stages. Those are the rising stage, support or maintenance stage, and the recession or subsidence stage. In the rising stage, the positioning estimation of the versatile application would be expanded suddenly while, in the support stage, the positioning estimation of portable would be kept up without corruption by giving profitable administrations to theclients. In the retreat stage, the positioning quality would be corruptedallofasuddenfromamoreelevatedamounttothe lower level. From this ranking investigation, we can anticipatethefakebyfindingthesuddenpositioningrisingor subsidencestage.

Rating evidences are also focused on to observe its increase or decrease anonymously. This can be doneto uplift the reputation of the apps and hence it is also considered as an important evidence. Overall, review evidence is vital and the key to determining the nature of an application. This can be done by taking into consideration of the various words present in the dataset.

The reviews go under a series of processes such as cleaning of the data, pre-processing them as stemming algorithms, using n-gram dataset to determine their polarity and rate them accordingly. With this N-gram dataset, we can split the required words (such as good or bad) from the other review words and each of the words are given a specified numerical value. Combining the values and taking an average with the original rating will help in determining the vast difference of anonymous rating with that of the one resulted by the actual sentiment process.

Overall, the comments are split up as words and each of the word is checked with the stored singlekey and multikey (N-gram) in the database. If the users commented words are matched with the one in database, the score of the keywords areretrievedforfurthercalculations.Userscommentscoreas well as the admin one is recalculated and stored as the new rating for theapplication.

The above results are aggregated as evidence result. This is then given as the output to the users in determining the fraud application by their ratings and reviews from the processes. In order to determine the fraud of the application, the rating stored in the database which is inclusive of the users rating score is compared to that of google play store and app store. If there is a vast difference, the application is sent for review. With this, the application gets eliminated fromthestoreinordertopreventfurtheruserdownloadsand fake reviews beingposted.

D. Algorithm

The admin is allowed to add and create new applications along with the links to the actual app in the play or appstore. A set of data is collected for that specific application from both the stores and saved in the database from a specific period of time. The user is able to view, download, rate and review the applications that are posted by the admin. Several data pre-processing methods are used in order to clean the datawhichhasbeengivenbytheuser.Asinthearchitecture, itcanbelogicallyvisualizedwiththetokenization,stopword removal and stemming algorithms beingused.

Here the user’s comments and reviews along with the singlekey and multikey words stored in the database act as the input to the algorithm. Based on these inputs, we are able to determine and get the score as our desired output. We initializethescoreandtheflagaszero.Whichmeansthatthe initial review based rating is set to zero. This would be modified and changed as per the words that are containedin thedatabaseaskeys.Theflagisthatwhichisalmostequalto the count function. As and when the words are read, the flag issetto0or1.Itrepresentsthatthewordispresentandread. As the output, the score value is determined which then reflects it on the users rating. This new score value is the usersrating.ThisalgorithmcanbedescribedasinAlgorithm 1.

SCORE CALCULATION

- input1: user's comment/reviewgiven
- input2: Single and multikeyvalues
- output: Score based on thereview
- Initializescore=0,flag=0
- Select multikey, singlekey whereflag=0
- get the score of singlekey= enteredstring
- get score of multikey=enteredstring
- score=(singlekey score or multikeyscore)/2
- return scorevalue

Algorithm 1

IV. RESULT AND CONCLUSION

This paper had presented about determining fraud applications by using the concept of data mining and sentiment analysis. It was supported by the architecture diagram which briefed about the algorithm and processes which are implemented in the project. Data gets collected and stored in the database which is then evaluated with the supporting algorithms defined. This is a unique approach in which the evidences are aggregated and confined into a single result. The proposed framework is scalable and canbe extended to other domain generated evidences for the rankingfrauddetection.Theexperimentalresultsshowedthe effectiveness of the proposed system, the scalability of detection algorithm as well as some regularity in the ranking fraudactivities.

REFERENCES

1. Daniel A. Keim, “Information Visualizing and Visual Data Mining” IEEE Trans. Visualization and Visual Data Mining, vol. 8,Jan-Mar 2002. *(references)*
2. FuzailMisarwala, KausarMukadam, and KiranBhowmick, “Applications of Data Mining in Fraud Detection”, vol. 32015.
3. 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.
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 Computing and Communication, vol. 6, February2018
7. Suleiman Y. Yerima, SakirSezer, Igor Muttik, “Android Malware Detection Using Parallel Machine Learning Classifiers”, 8th International Conference on Next Generation Mobile Applications, Services and Technologies,Sept.2014.
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 &Science,Vol. 4, October2018

Retrieval Number: B11070782S319/19©BEIESP
DOI : 10.35940/ijrte.B1107.0782S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

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. Monika Pandey, Prof. TriptiSharma,“Fraud App Detection using Fuzzy Logic Model Based on Sentiment of Reviews”, International Research Journal of Engineering and Technology,Vol. 5, Sep2018.
14. MahmudurRahman, MizanurRahman, BogdanCarbunar, and DuenHorngChau,“Search Rank Fraud and Malware Detection in Google Play”, IEEE Transactions on Knowledge and Data Engineering,Vol. 29, June2017.
15. 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.
16. (Online) Available:https://www.makeuseof.com/tag/who-invented- the-first-computer/.
17. (Online) Available: https://lifehacker.com/why-you-shouldnt-trust- app-store-reviews-and-what-to-1515379780
18. 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.
19. Dr.R.Subhashini and Akila G,"Valence arousal similarity based recommendation services ", IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2015.
20. Gladence, L. Mary, M. Karthi, and V. Maria Anu. &quot;A statistical comparison of logistic regression and different Bayes classification methods for machine learning.&quot; ARPN Journal of Engineering and Applied Sciences 10, no. 14 (2015): 5947-5953.