A Project Report

on

Fraud App Detection App Using Sentiment Analysis

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of

Dr. Kuldeep Singh Kaswan

Professor

Department of Computer Science and Engineering

Submitted By

18SCSE1010600 - TARUN GARG

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA, INDIA

DECEMBER - 2021

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Candidate's Declaration

I hereby declare that the work presented in this report entitled "Fraud App Detection Using Sentiment Analysis" in partial fulfilment of the requirement for the award of the degree of Bachelor of computer science and Engineering/Information Technology. Submitted in the Department of Computer Science and Engineering, Galgotias University is an authentic record of my own work carried out over a period from August to November 2021 under the supervision of Prof. (Dr) Kuldeep Singh Kaswan.

(Student Signature)

Tarun Garg

18SCSE1010800

This is certified that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)

Dr. Kuldeep Singh Kaswan

Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **18SCSE1010600 - TARUN GARG** has been held on __and his work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:

Place:

ACKNOWLEDGEMENT

Primarily We might thank God for having the ability to finish this mission with success .Then we would love to thank my mission manual Dr. Kuldeep Singh Kaswan ,whose treasured steerage has been those that helped me patch this mission and make it complete evidence success, his pointers and his commands has served because the foremost contributor in the direction of the finishing touch of the mission. Then We would love to thank our mother and father and friends who've helped me with their treasured pointers and steerage has been beneficial in diverse levels of the finishing touch of the mission.

Abstract

Introduction-

With the growth of technology, there is an increase in mobiles use There has been significant growth in the development of various mobile applications for many popular platforms like in Android and IOS. The most important role played by customers status, ratings and reviews for that particular app when the download happens. when the download happens. This could be a way for Developers find their weaknesses and improve on the development of a new one of taking into account the needs of the people. As cell phones have become a popular necessity, so have they it is important that suspicious applications be marked as fraud to be seen by store users. It will be so it is difficult for a user to decide which ideas scroll through either the scales they see as scam or real for their benefit. Thus, we propose a plan will identify those fraudulent apps on Google Play or App finally give a complete overview of the discovery of fraud system.

Proposal of Work

We judge the security and authenticity of each application are based solely on updates what is said in each program. It is therefore a necessity keep track and upgrade the system to make sure applications are available are true or not. The aim is to improve the system to detect fraudulent applications before downloading the user through emotional analysis and data mining.

Emotional analysis is help in determining emotional tones after word displayed online. This approach helps in social awareness media also helps to get a brief overview of public opinion a about certain issues. User may not always receive relevant or accurate updates about an online product. We can look at the user sympathetic ideas for much application. Reviews are possible be false or true. Rate analysis and reviews together including user and management comments, we can see whether the app is real or not. Emotional analysis is used and data mining, this the machine can read and analyze emotions, feelings about updates and other texts. The update fraud is one of the most important aspects of app ratings fraud.

Through emotional analysis and data mining, analysis reviews and comments can help you find the right one Android and IOS platforms application

Work Importance-

In today's market there are thousands of fake apps on play store and apple store the result of this fake apps that the user installs it in same of the fake apps can cause malware in the Device it can also lead to privacy. These fake apps can steal data which is relevant to the User the data for the sometime includes their bank details their passwords and social media Accounts so our app will ensure that the user gets saved from this fake apps and will protect their data and password.

Table of Content

Title

Candidates Declaration

Acknowledgement

Abstract
List of Figures and Tables

Chapter 1 Introduction
Introduction of project
Formulation of Problem
Tool and Technology Used

Chapter 2 Literature Survey/Project Design
Literature Survey
System Analysis
Proposed Methodology
Modules

Chapter 3 Functionality and Concept of Project
Objectives
System Diagram
Proposed work
Existing System
Score Calculation

Chapter 4 Results and Discussion
Result
Conclusion
Future Scope
Reference

List of Figure

Figure No.	Figure Name	Page Number
1	Architectural Diagram	4
2	Fig.2- Ranking based Evidence	5
3	Fig.3- Rating based Evidences	6
4	Fig 4- Review based Evidences	6
5	Fig5-Evidenceaggregation	7

Chapter - 1

Introduction:-

Sentiment is an emotion or attitude prompted by the feelings of the customer. Sentiment analysis is also referred to as opinion mining, as opinions are collected from customer is mined to reveal the rating of the app. The process of Sentiment analysis comes under machine learning. [1] Information is gathered and is analyzed to determine the sentiment about the information such as negative or positive sentiment. Before purchasing the app people always enquire about the opinion of the app by the other users. [2] The process of Sentiment analysis uses natural language processing (NLP) to collect and examine the opinion or sentiment of the sentence. It is popular as many people prefer to take some advice from the users. As the amount of opinions in the form of reviews, blogs, etc. are increasing continuously, it is beyond the control of manual techniques to analyze huge amount of reviews & to aggregate them to a efficient decision. Sentiment analysis performs these tasks into automated processes with less user support. [3] It is not always possible to have a one technique to fit in all solution because different types of sentences express sentiments/opinions in different ways. Sentiment words (also called as opinion words) (e.g., great, beautiful, bad, etc) cannot distinguish an opinion sentence from a non-opinion one. A conditional sentence may contain many sentiment words or sentences, but express no opinion. The type of sentences, i.e., conditional sentences, it have some unique characteristics which make it hard to determine the orientation of sentiments on topics/features in such of the sentences. By sentiment orientation, we mean positive, negative or neutral opinions. Conditional sentences are sentences which describe implications or hypothetical situations & their consequences. In English language, a variety of conditional connectives can be used to form these sentences. A conditional sentence contains two clauses: the condition clause and the consequent clause, that are dependent on each other. Their relationship has significant implications on whether the sentence describes an opinion. [4] As there are more than millions of apps on the App store, there is many competition between apps to be on top of the leader board on the basis of popularity. As leader board is the most important way for promoting apps. The higher rank on the leader board leads to huge number of downloads & million doll or of profit. Apps give advertisement to promote their

apps on the leader board. Many apps use fraudulent means to boost their ranking on the leader board of the App store. There are various means to increase downloads & ranking of the app which is done by "bot farms" or "human water armies", human water armies are a group of internet ghostwriters who are paid to post fake reviews. The app is said to fraud on the basis of 3 parameters: Ranking, Rating & Review of the app. In ranking based we check the historical ranking of the app, there are 3 different ranking phases, rising phase, maintaining phases & recession phase. The apps ranking rising to peak position on leader board (ie. rising phase), to keep at the peak position on the leader board (ie. maintaining phase), & finally decreasing till the end of event (ie. recession phase). The reviews are taken from the dataset and are converted into tokens on which sentiment analysis is performed.

The most important role played by customers quality, ratings and reviews of that particular app what happens to download. Not that, sometimes developers are misleading recognition of their applications or malicious use them as a malware distribution platform throughout. Occasionally, it is just an improvement for engineers they often hire teams of workers who commit fraud by sharing and providing false opinions and estimates over application. This is known as crowd turfing. It is therefore important to make sure that before installing app, users are provided with the right and true comments to avoid something wrong. In this case, I an automated solution is needed to win again systematically analyze various ideas and measurements provided for each application. It will be so it is difficult for a user to decide to comment on what they are saying scroll through even if the scales they see are fraudulent or true for their benefit. Thus, we are proposing a system of that will detect malicious applications on Google Play or App end by giving a complete overview of fraud detection by scale system. By considering data mining and the emotional analysis, we can get a higher probability to get real reviews with us suggest a program that takes reviews from registered users with one or more products and test them as a positive or negative rating. This can be helpful as well determine the application for fraud and ensure mobile security well we check in three forms proof based, based ratings, and model-based reviews are three-dimensional combinations with statistics hypotheses. Regardless, evidence based on suspension may be affected by the status of the application developer and others real marketing efforts “as a set time laying down “.

In this project we mainly focus to get the review by the users is genuine. users can use the app by just signing up and write their reviews in the review section and write the name of the app in the app name section. and now admin can check their reviews in the review section in the review section it shows all the reviews which were written by the different - different reviewer. and in the chart section they show all the app rating scale based on the reviews. And reviewer can check the review that which type of review they written they will check in the dataset.

Developers have developed a ranking fraud detection system for mobile Apps. Specifically, we show that ranking fraud happened in the leading sessions and provided a method for mining leading sessions for each App from its historical ranking records. Then, we identify ranking based evidences and rating based evidences for detecting ranking fraud. Moreover, we proposed an optimization based aggregation method to integrate all the evidences for evaluating the credibility of leading sessions from mobile Apps. An unique perspective of this approach is that all the evidences can be modeled by statistical hypothesis tests, thus it is easy to be extended with other evidences from domain knowledge to detect ranking fraud. Finally, we validate the proposed system with extensive experiments on the real-world App data collected from the App store. Experimental results showed the effectiveness of the proposed approach.[15] The main objective is fraud application detection using fuzzy logic to differentiate the actual fraud apps. The proposed system perform classification of apps & detect their group whether they belong to good, bad, neutral, very good, very bad. Different class value & threshold value gives different results of accuracy of time required for execution.[16] Sentiment Analysis is major task of NLP (natural language processing). Data used as input are online app reviews. The objective content from the sentences are removed and subjective content is extracted. The subjective content consists of sentiment sentences. In NLP, part-of-speech (POS) taggers are developed to classify words based on POS. Adjective and verbs convey opposite sentiment with help of negative prefixes. Sentiment score is computed for all sentiment tokens.

Information is gathered and is analyzed to determine the sentiment about the information such as negative or positive sentiment. Before purchasing the app people always enquire about the opinion of the app by the other users. The process

of Sentiment analysis uses natural language processing (NLP) to collect and examine the opinion or sentiment of the sentence. It is popular as many people prefer to take some advice from the users. As the amount of opinions in the form of reviews, blogs, etc. are increasing continuously, it is beyond the control of manual techniques to analyze huge amount of reviews & to aggregate them to an efficient decision. Sentiment analysis performs these tasks into automated processes with less user support. It is not always possible to have a one technique to fit in all solution because different types of sentences express sentiments /opinions in different ways. Sentiment words (also called as opinion words) (e.g., great, beautiful, bad, etc.) cannot distinguish an opinion sentence from anon-opinion one. A conditional sentence may contain many sentiment words or sentences, but express no opinion. The type of sentences, i.e., conditional sentences, it has some unique characteristics which make it hard to determine the orientation of sentiments on topics/features in such of the sentences. By sentiment orientation, we mean positive, negative or neutral opinions. Conditional sentences are sentences which describe implications or hypothetical situations & their consequences. In English language, a variety of conditional connectives can be used to form these sentences. A conditional sentence contains two clauses: the condition clause and the consequent clause, that are dependent on each other. Their relationship has significant implications on whether the sentence describes an opinion. As there are more than millions of apps on the App store, there is many competitions between apps to be on top of the leader board on the basis of popularity. As leader board is the most important way for promoting apps. The higher rank on the leader board leads to huge number of downloads & million doll or of profit. Apps give advertisement to promote their apps on the leader board. Many apps use fraudulent means to boost their ranking on the leader board of the App store. There are various means to increase downloads & ranking of the app which is done by "bot farms" or "human water armies", human water armies are a group of internet ghostwriters who are paid to post fake reviews. The app is said to fraud on the basis of 3 parameters: Ranking, Rating & Review of the app. In ranking based we check the historical ranking of the app, there are 3 different ranking phases, rising phase, maintaining phases & recession phase. The apps ranking rising to peak position on leader board (i.e. rising phase), to keep at the peak position on the leader board (i.e. maintaining phase), & finally decreasing till the end of event (i.e. recession phase). There views are taken from the dataset and are converted into tokens on which sentiment analysis is performed. Volume No: 6(2021), Issue No: 12(May) www.ijracse.com Page 11 1.2 Sentiment Analysis Techniques 1.2.1

Machine Learning Machine learning based Sentiment Analysis or arrangement should be possible in two different ways: Sentiment Analysis by utilizing directed machine learning strategies and Sentiment Analysis by utilizing unsupervised machine learning procedures. (a) Supervised Machine Learning In Supervised Machine learning procedures, two sorts of informational collections are required: preparing informational index and test informational collection. A programmed classifier takes in the grouping variables of the report from the preparation set and the exactness in order can be assessed utilizing the test set. Various machine learning calculations are accessible that can be utilized extremely well to characterize the records. The machine learning calculations like Support Vector Machine (SVM), Naive Bayes(NB) and greatest entropy(ME) are utilized effectively in numerous examinations and they performed well in the feeling characterization. (b) Unsupervised Machine Learning Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. It is a semantic orientation way to deal with opinion mining in which sentiment polarity of highlights show in the given record are controlled by contrasting these highlights and semantic lexicons. Semantic dictionary contains arrangements of words whose sentiment orientation is resolved.

1.2 Sentiment Analysis Techniques 1.2.1 Machine Learning Machine learning based Sentiment Analysis or arrangement should be possible in two different ways: Sentiment Analysis by utilizing directed machine learning strategies and Sentiment Analysis by utilizing unsupervised machine learning procedures. (a) Supervised Machine Learning In Supervised Machine learning procedures, two sorts of informational collections are required: preparing informational index and test informational collection. A programmed classifier takes in the grouping variables of the report from the preparation set and the exactness in order can be assessed utilizing the test set. Various machine learning calculations are accessible that can be utilized extremely well to characterize the records. The machine learning calculations like Support Vector Machine (SVM), Naive Bayes(NB) and greatest entropy(ME) are utilized effectively in numerous examinations and they performed well in the feeling characterization. (b) Unsupervised Machine Learning Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. It is a semantic orientation way to deal with opinion mining in which sentiment polarity of highlights show in the given record are controlled by contrasting these highlights and semantic lexicons. Semantic dictionary contains arrangements of words whose sentiment orientation is resolved as of now.

It arranges the archive by conglome rating the sentiment orientation of all assessment words display in the record, reports with more positive word lexiconsis characterized as positive document and the documents with more negative word lexicons is classified as negative document. Hybrid Technique: A few researchers combined the supervised machine learning and lexicon-based techniques jointly to enhance sentiment classification performance. They considered both general reason lexicon and domain specific lexicon for identifying polarity orientation of sentiment words and feed these lexicons into supervised learning algorithm, SVM. They found that general purpose lexicon performed very poor while domain specific lexicon performed very well. The system classified the sentiment in two steps: First the classifier is trained to predict the aspects and In Next the classifier is trained to predict the sentiments related to the aspects collected in step. Their system yielded around 66.8%accuracy. 1.3 Lexical Information Understanding human emotions (also called sentiment analysis) is a hard job for a machine, for which the computational intelligence technique may offer enhanced results. Regularly, semantic articulations and additionally paralinguistic includes in talked dialects (e.g., pitch, loudness, tempo, etc.) reveal the sentiments or emotional states of individuals. Prior research studies have developed sentiment lexicons using a Volume No: 6(2021), Issue No: 12(May) www.ijracse.com Page 12 dictionary technique and a corpus technique. Social media has changed the world. It has become an everyday part of our lives. Many people are nowadays active on several popular social networks such as Face book, twitter, Instagram, etc. They share photos and posts on their daily life and experiences such as their food, their clothes, and their trips. Some people are more active on social networks, while others are less so.

Most of us use android and IOS Mobiles these days and also uses the play store or app store capability normally. Both the stores provide great number of application but unluckily few of those applications are fraud. Such applications dose damage to phone and also may be data thefts. Hence, such applications must be marked, so that they will be identifiable for store users. So we are proposing a web application which will process the information, comments and the review of the application. So it will be easier to decide which application is fraud or not. Multiple application can be processed at a time with the web application. Also User cannot always get correct or true reviews about the product on internet. So rating/comments will be judged by the admin and it would be easy for admin to predict the application as

Genuine or Fraud. Advantages The proposed framework is scalable and can be extended with other domain generated evidences for ranking fraud detection. Experimental results show the effectiveness of the proposed system, the scalability of the detection algorithm as well as some regularity of ranking fraud activities. To the best of our knowledge, them is no existing benchmark to decide which leading sessions or Apps really contain ranking fraud. Thus, we develop four intuitive baselines and invite five human evaluators to validate.

2. System Analysis

2.1 Existing System A general approach to protect the data confidentiality is to encrypt the data before outsourcing. Searchable encryption schemes enable the client to store the encrypted data to the cloud and execute keyword search over cipher text domain. So far, abundant works have been proposed under different threat models to achieve various search functionality, such as single keyword search, similarity search, multi-keyword Boolean search, ranked search, multi-keyword ranked search, etc. Among them, multi-keyword ranked search achieves more and more attention for its practical applicability. Recently, some dynamic schemes have been proposed to support inserting and deleting operations on document collection. These are significant works as it is highly possible that the data owners need to update their data on the cloud server. 2.1.1 Disadvantages Of Existing System Huge cost in terms of data usability. For example, the existing techniques on keywordbased information retrieval, which are widely used on the plaintext data, cannot be directly applied on the encrypted data. Downloading all the data from the cloud and decrypt locally is obviously impractical. Existing System methods not practical due to their high computational overhead for both the cloud sever and user.

2.2 Proposed System This paper proposes a secure tree-based search scheme over the encrypted cloud data, which supports multi-keyword ranked search and dynamic operation on the document collection. Specifically, the vector space model and the widely-used “term frequency (TF) × inverse document frequency (IDF)” model are combined in the index construction and query generation to provide multikey word ranked search. In order to obtain high search efficiency, we construct a tree-based index structure and propose a “Greedy Depthfirst Search” algorithm based on this index tree. The secure KNN algorithm is utilized to encrypt the index and query vectors, and meanwhile ensure accurate relevance score calculation between encrypted index and query vectors. To resist different

attacks in different threat models, we construct two secure search schemes: the basic dynamic multi-keyword ranked search (BDMRS) scheme in the known ciphertext model, and the enhanced dynamic multi-keyword ranked search (EDMRS) scheme in the known background model.

1.2 Sentiment Analysis Techniques 1.2.1 Machine Learning Machine learning based Sentiment

Problem Definition:

1. Changing fraud patterns over time - This is very difficult to deal with as fraudsters are always looking to find new and innovative ways to go around the plans to commit this act. It is therefore very important that in-depth learning models are updated with advanced patterns for recognition. This results in a decrease in the efficiency and effectiveness of the model. Machine learning is the models therefore need to constantly update or fail their goals.

2. Class Inequality - Only a small percentage of customers have fraudulent intentions. As a result, there are inequalities in classifying fraud detection models (which often classify fraudulent or non-fraudulent) making it difficult to enforce. At the root of this challenge is poor user behavior towards real customers, as catching scammers often involves a decline in certain legitimate activities.

3. Model Definitions - This limitation is associated with the definition of interpretation as models often give points that indicate whether the work may be fraudulent or not - without explaining .

4. Feature construction may be time consuming - Mathematicians may need a lot of time to create a comprehensive set that delays the process of detecting fraud.

Apk file of mobile application is uploaded on the web application. APK parser is used to extract information about the application such as reviews, ratings and historical record. Natural Language Processing is used to perform sentiment analysis on the reviews. By applying rule for detection of fraud application, it generates the graph results. If the rating count is greater than 3 then it is considered as a positive result. And if the rating count is less than 3 then it is considered as a

negative result. Methodology used are cloud stack, data mining and NLP. [6] Application reviews are extracted and converted into tokens. Tokenization is process of converting a stream of text into words, phrases, symbols known as tokens. This tokens are the input for pre processing. After preprocessing of reviews system determine the user emotions. Positive reviews add 1 to positive score and negative review adds 1 to negative score. With this it will determine score of every review and confirm whether the application is real or fake. [7] In this paper two methods of sentiment analysis are compared. Lexicon based approach and machine learning approach. Lexicon based approach deals with searching the sentiment words form the sentence and comparing with existing list of words, it has two branches dictionary and corpus based approach. Lexicon–based approach does not require training set whereas naïve bayes requires training set Lexicon–based method is accurate than Naïve bayes classifier when sentence is processed completely with training set data. [8] User reviews are collected using open source scrapping tools and stored in my SQL database. Titles and comments are extracted from stored dataset. Collocation finding algorithm provided by NLTK toolkit is used for extraction of features from user reviews. user sentiments are extracted about the identified features and given them a general score across all reviews. Finally topic modeling techniques are used to group fine grained features into more meaningful high-level features. [9] In this paper The Tweets Sentiment Analysis Model analyses tweets data. It can identify positive, negative or neutral sentiments and measure intensity of positive/negative opinions in regard to any category. The framework of the TSAM consists of three modules: Feature selection module that extracts the relevant words from each sentence Sentiment identification module that associates expressed opinions with each relevant entity in each sentence level. Sentiment aggregation and scoring module calculates the sentiment scores for each entity. [10] Google API calculation approach is used to calculate the rank of the applications using Calculation algorithm where they take application ratings from play store and calculate the ranks using the calculations. [11] Feature extraction in sentiment analysis is an emergent research field so in this paper we have concentrated on related work performed to identify directions for future work. There are many feature selection techniques, NLP based, Machine learning or clustering based, Statistical, Hybrid, are discussed. Features are categorized as syntactic, semantic, lexico-structural, implicit, explicit and frequent, making it easy for the future researchers to work on. Different pre-processing modules like POS tagging, stop word removal, stemming and lemmatization are discussed. Finally we conclude that feature space reduction, redundancy removal and evaluating

performance of hybrid methods of feature selection can be the future direction of research work for all researchers in the field of feature extraction in sentiment analysis.

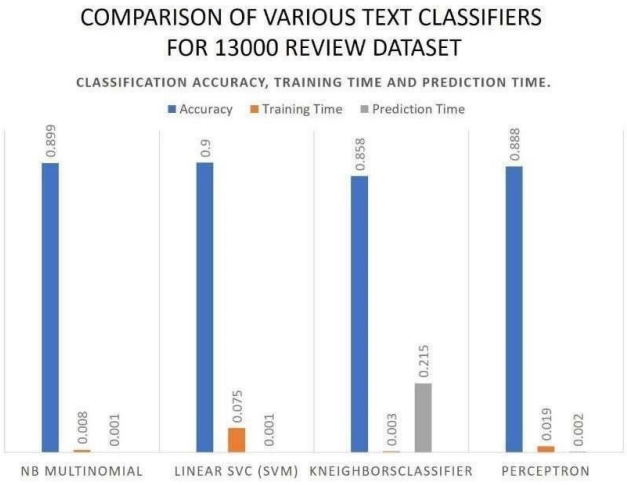
Tool and Technology Used:

We have compared methods based on four different classification methods:

1. Multinomial naïve bayes
2. linear SVM
3. K-neighbor
4. Perception

Sr no Method Accuracy Training time Prediction time 1 Multinomial naïve bayes 89.9% 0.008 sec 0.001 sec 2 linear SVM 90% 0.075 sec 0.001 sec 3 K-neighbor 85.8% 0.003 sec 0.003 sec 4 Perception 88.8% 0.019 sec 0.002 sec Table2.2 : Comparison table As seen in above table multinomial naïve bayes has an accuracy of 89.9% and less computing time hence it is more preferred over other methods. Our fraud application detection method uses it as classification method.

In this report we ensure that the users can get the genuine review for this we create the signing the page for the users and write their genuine reviews in the review section admin can also ensure that the reviews are written correctly or it is fake or



real. Reviewer can write the review with the help of the data set. in the particular app the different- different review are there. so,

PROPOSED METHODOLOGY

DATA COLLECTION DETAILS Data collection is an important part of Machine Learning. Data collection is the process of gathering and measuring information for different available sources. Machine Learning requires a huge set of data having multiple attributes, to be able to classify some input parameters more accurately. Data collection is the important aspect that makes the algorithm training possible. It has been observed that greater number of attributes yields a better result. Training data for fraud app detection is obtained from Training data Training dataset was used for training the algorithm so that algorithm learns and produce results. Training dataset consist of 13000 entries (reviews, sentiment value). Training dataset consists 50% of positive and 50% of negative reviews. Testing dataset: Testing dataset was used for evaluating the model/algorithm with trained dataset. Testing dataset is real time dataset which is extracted from google playstore. DATA PREPROCESSING The process of converting data to something a computer can understand is called Preprocessing. The dataset which is obtained in data collection is not in the form which can be used by the classifier. Various Data preprocessing and feature extraction techniques must be performed on the dataset to make it suitable for generation of classification model. The python library Pandas is used to perform the preprocessing techniques on the dataset. Preprocessing steps are: Tokenization Tokenization basically refers to splitting up a large body of text into smaller lines, words or even creating words for a non-English language. Various tokenization functions are inbuilt into nltk module itself. Stopwords Removal Stop Word Removal is a process of filtering out useless data. In NLP, useless words are referred to as stopwords. Lowercase conversion In this all the upper case letters are converted to lower case. Tfidf Vectorizer The Tfidf Vectorizer will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow you to encode new documents. PROPOSED RESEARCH DESIGN The proposed approach for the system can be carried out by using corpus based and Naïve Bayes based approach to detect fraud application. First the dataset is prepared so that it can be used for the classifier. The dataset is first stored in a data structure dataframe which can be made by the pandas library. By using the tfidfVectorizer function, various features are extracted

3.2 Objective of The System To design a system which may detect fake apps by considering different evidences indicating their true behavior. To find apps are real or not. To increase the classification accuracy of a system.

3.3 Functional Requirements In Software engineering, a functional requirement defines a function of a software system or its components. A function is described as set of inputs, the behavior and outputs. Functional requirements may be calculations, technical details, data manipulation and processing and other specific functionality that define what a system is supposed to accomplish. Behavioral requirements describing all the cases where the system uses the functional requirements are captured in use cases. Functional requirements are supported by non-functional requirements which impose constants on the design or implementation.

3.3.1 Functional Requirements For This Project The functional requirements describe the interactions between the system and its environment independent of its implementations.

MODULE 1: Data Collection Details Data collection is an important part of Machine Learning. Data collection is the process of gathering and measuring information for different available sources. Machine Learning requires a huge set of data having multiple attributes, to be able to classify some input parameters more accurately. Data collection is the important aspect that makes the algorithm training possible. It has been observed that greater number of attributes yields a better result. Training data for fraud app detection is obtained from Training data Training dataset was used for training the algorithm so that algorithm learns and produce results. Training dataset consist of13000 entries (reviews, sentiment value). Training data set consists 50% of positive and 50% of negative reviews. Testing dataset: Testing dataset was used for evaluating the model/algorithm with trained dataset. Testing dataset is real time data set which is extracted from google play store.

MODULE 2: Data Preprocessing The process of converting data to something a computer can understand is called Preprocessing. The dataset which is obtained in data collection is not in the form which can be used by the classifier. Various Data preprocessing and feature extraction techniques must be performed on the dataset to make it suitable for generation of classification model. The python library Pandas is used toper form the preprocessing techniques on the dataset. Preprocessing steps are: Tokenization Tokenization basically refers to splitting up

a large body of text in to smaller lines, words or even creating words for a non-English language. Various to kenization functions are in built into module itself.

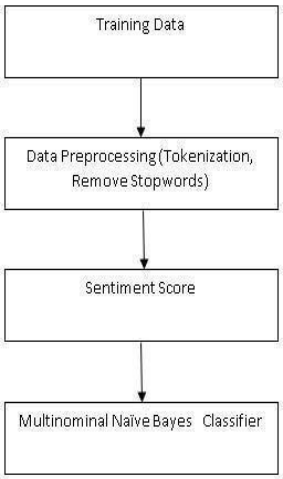
MODULE 3: Stop words Removal Stop Word Removal is a process of filtering out useless data. In NLP, useless words are referred to ass top words. Lower case conversion In this all the upper-case letters are converted to lower case. Tfidf Vectorizer The Tfidf Vectorizer will to kenize documents, learn the vocabulary and inverse document frequency weightings, and allow you to encode new documents.

3.4 Non-Functonal Requirements In System engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system rather than specific behavior. In general, functional requirements define what a system is supposed to do whereas non-functional requirements define how a system is supposed to be.

3.5 System Requirements

3.5.1 Software Requirements Operating system: - Windows XP. Coding Language: J2EE Data Base: MYSQL 3.5.2 Hardware Requirements System: Pentium IV 2.4 GHz. Hard Disk: 40 GB. Floppy Drive: 1.44 Mb. Monitor: 15 VGA Colour. Mouse: Logitech. Ram: 512 Mb. 3.6 Software Development Environment

3.6.1 An Introduction to Python Python is a popular object-oriented programming language having the capabilities of high-level programming language. Its easy to learn syntax and portability capability makes it popular these days. The chapter describes about the software tool that is used in our project. Python was developed by Guido van Rossum at Stichting Mathematisch Centrum in the Netherlands. It was written as the successor of programming language named ‘ABC’. It’s first version was released in 1991. The name Python was picked by Guido van Rossum from a TV show named Monty Python’s Flying Circus. It is an open source programming language which means that we can freely download it and use it to develop programs.



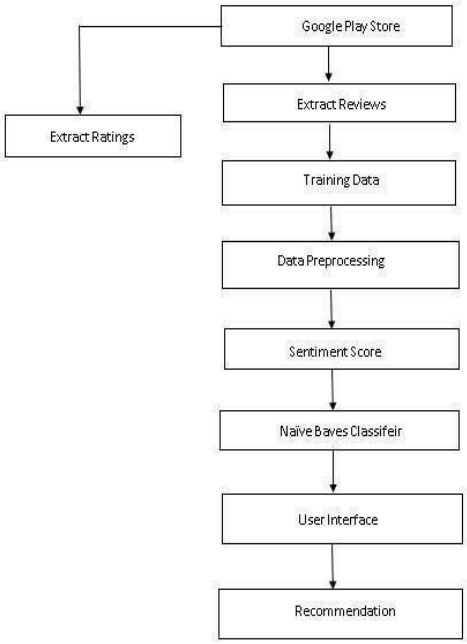
they will categorize the review in to three categories.

1) Positive

2) Negative

3) Neutral

And in the chart section they show graph of the rating of apps. the graph is based on the scale of review versus app. In this app we use the python programming and use flask for storing the files and and in the database in my SQL to store the data in



the dataset.

4. System Design 4.1 Introduction The design of a system is essentially a blueprint or a plan for a solution for the system. Here we consider a system to be a set of components with clearly defined behavior that interacts with each other in a fixed defined manner to produce some behavior or services for its environment.

4.1.1 Input Design The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things: What data should be given as input? How the data should be arranged or coded? The dialog to guide the operating personnel in providing input. Methods for preparing input validations and steps to follow when error occur.

4.1.2 Objectives

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus, the objective of input design is to create an input layout that is easy to follow.

4.1.3 Output Design A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output

design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system. The output form of an information system should accomplish one or more of the following objectives. o Convey information about past activities, current status or projections of the Future. o Signal important events, opportunities, problems, or warnings. o Trigger an action. o Confirm an action.

4.2 Database Design A general theme begins a database is to handle information as an integrated whole. A database is a collection of inter-related data stored with minimum redundancy to server many users quickly and efficiently. The general objective is to make information access easy, quick, expensive and flexible for the user. In database design several specific objectives are consider:

4.2.1 Control Redundancy Redundant data occupies space and therefore, is wasteful. If versions of the same data are in different phase of updating, a system often gives conflicting information. A unique aspect of database design is storing data only once, which controls redundancy and improves system performance.

4.2.2 Data Independence An important database objective is changing hardware and storage procedures for adding raw new data without having to rewrite application programs.

4.2.3 Accuracy and Integrity The accuracy and database ensure the data quality content remain constant. Integrity controls detects data inaccuracy where occur.

4.2.4 Privacy and Security For the data to remain private, security measures must be taken to an unauthorized access. Database security means that data are protected from various forms of destructions. Uses must be positively identifies and actions monitored. Managing the database require a Database Administrator (DBA) whose

key functions are to be managing data activities, The database structure and the DBMS. In addition, a managerial background the DBA needs a technical knowledge to deal with database design.

. Testing 6.1 Introduction The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement. 6.1.1 Testing Objectives All field entries must work properly. Pages must be activated from the identified link. The entry screen, messages and responses must not be delayed. Testing cannot show the absence of defects, it can only show that software errors are present. Once code has been generated the testing begins. The purpose of testing is more than just debugging and detecting of bugs. Testing is usually performing for the following: o For improving and assuring software quality. o For estimation reliability. o For verification and validation. The objection that should kept in mind while testing is being executed as follows: o It should be easily predictable. o It should be fixed. o It should follow certain constraints a rule. 6.1.2 Testing Approaches Field testing will be performed manually and functional tests will be written in detail.

The different types of Tests are as follows:

1) Unit Testing

2) Integration Testing

3) Functional Test

4) System Test

5) White Box Testing

6) Black Box Testing

7) Acceptance Testing

6.2 Unit Testing Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce

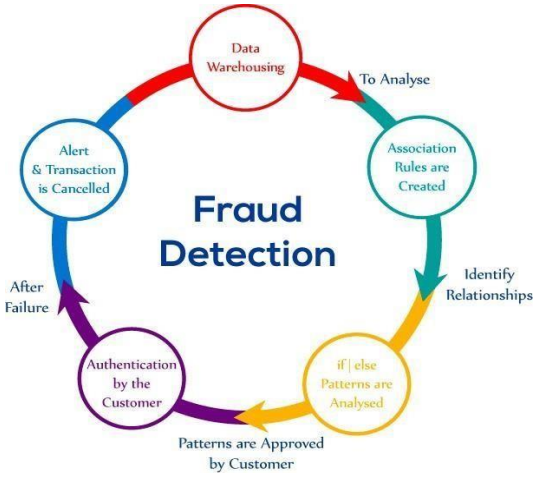
valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results. Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases. Integration Testing Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned .

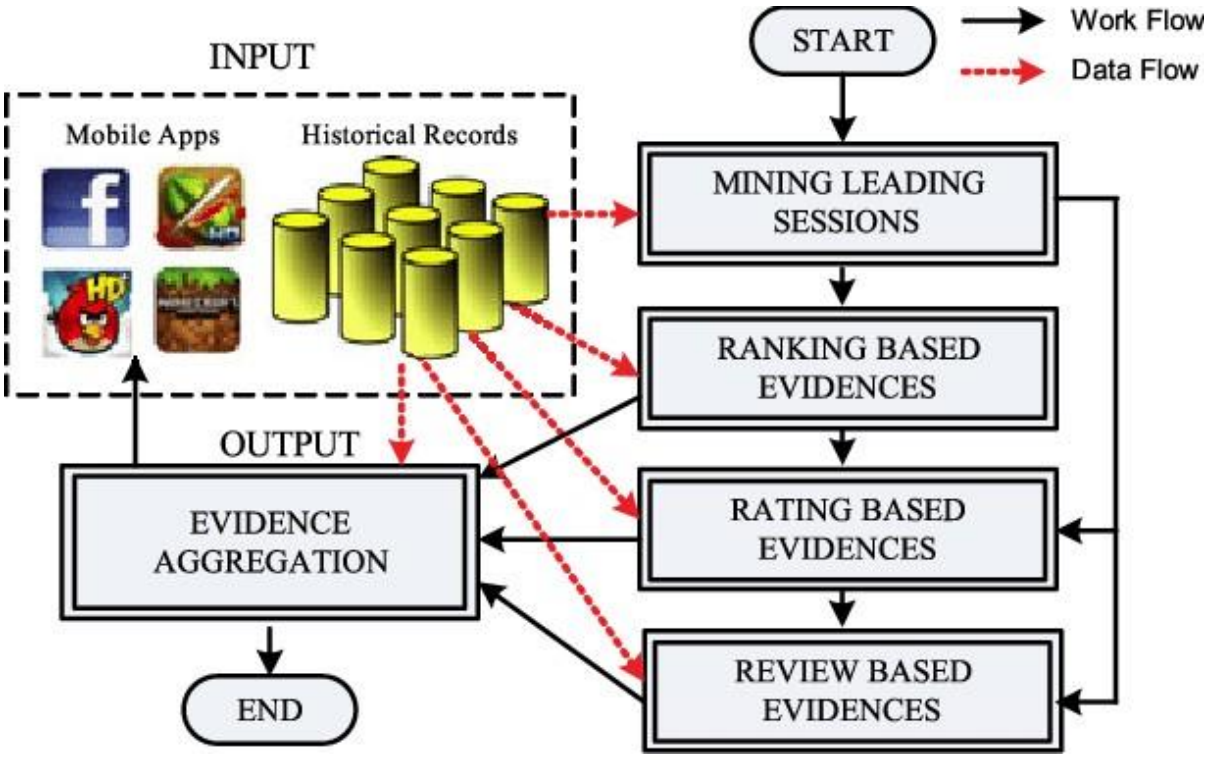
with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components. Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error. 6.4 Functional Test Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items: Valid Input: identified classes of valid input must be accepted. Invalid Input: identified classes of invalid input must be rejected. Functions: identified functions must be exercised. Output: identified classes of application outputs must be exercised. Systems/Procedures: interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

6.5 System Test System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable

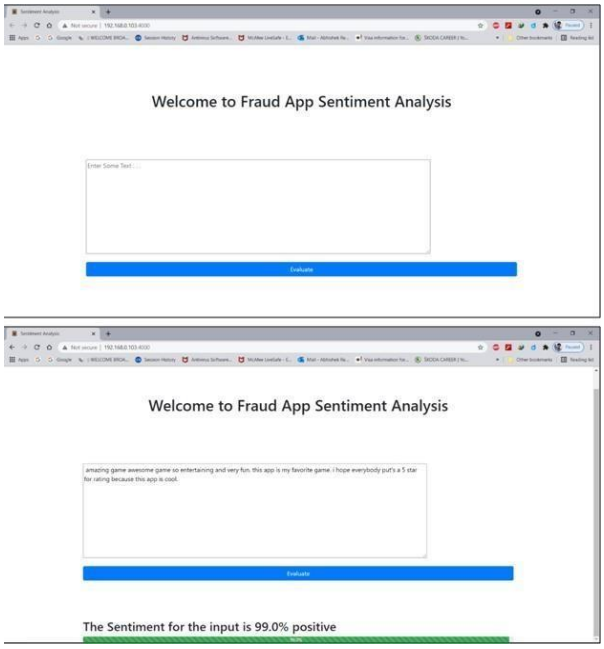
results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points. 6.6 White Box Testing White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

System Diagram:

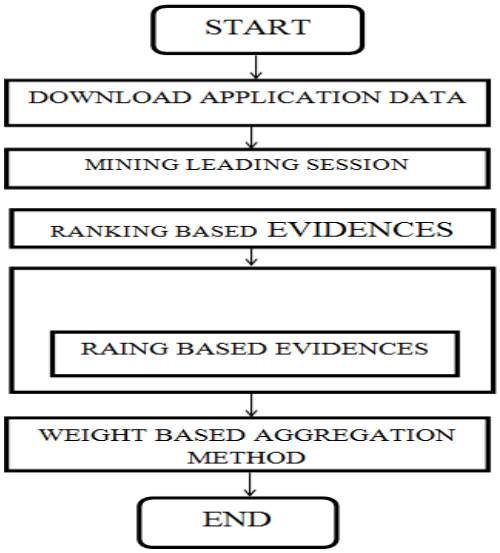


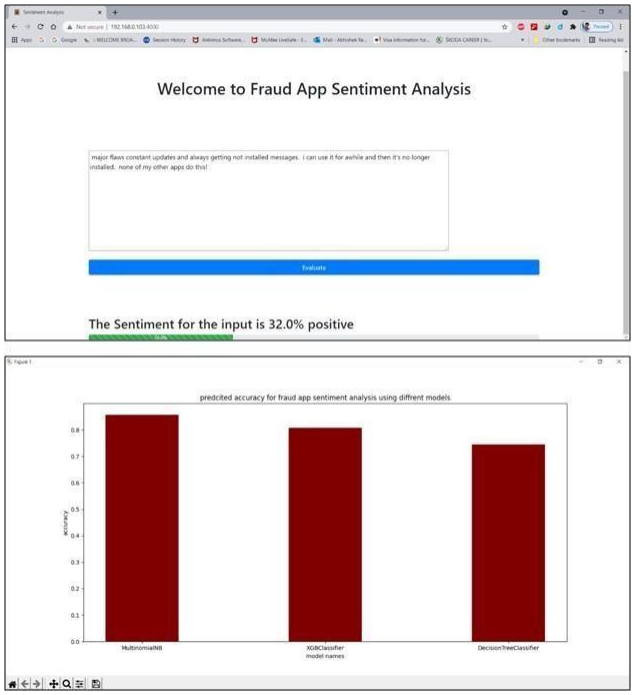

START
Work Flow
Data Flow
INPUT
Mobile Apps
Historical Records
MINING LEADING SESSIONS
RANKING BASED EVIDENCES
RATING BASED EVIDENCES
REVIEW BASED EVIDENCES
OUTPUT
EVIDENCE AGGREGATION
END

Proposed work:



Description From the above figure, the window clearly tells that the evaluate of given review produces 99.0% positive and High accuracy rate.





Description From the above figure, we can see the predicted accuracy for fraud app sentiment analysis using different models i.e. Multinomial NB, XGB Classifier, and Decision Tree Classifier. Among those three models Multinomial NB produced high accuracy.

8. Conclusion In this project, we have conducted a survey regarding different methodologies used in reviewing the status of application and predicting whether it is fraud or not. Our proposed methodology deals with sentiment analysis which has an advantage over the other methods due to fact that lexicon-based analysis is more accurate and faster than other approaches. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you would like less training data. It is a fast algorithm for classification problems. It is an honest fit real time prediction, multiclass

prediction, recommendation system, text classification and sentiment analysis use cases. Naive Bayes Algorithm are often built using Gaussian, Multi nomial and binomiall distribution. it's very low computational cost.

intrigue and advancement since it has numerous handy applications. Since freely and secretly accessible data over the Internet is continually growing, countless communicating conclusions are accessible in audit locales, discussions, online journals, and web-based social networking. 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 ,the text is evaluated. To be particular, we are using text data mining which is also referred as text mining. From the texts which are extracted(reviews) it is easier to analyze words or a cluster of words that are used. 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 user details from the store. The admin add sane application 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 and events. 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 the clients. In the retreat stage, the positioning quality would be corrupt dall of a sudden from a more elevated amount to the lower level. From this ranking investigation, we can anticipate the fake by finding the sudden positioning rising or subsidence stage. 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 DETECTING FRAUD APPS USING SENTIMENT RESEARCH 583 Published By: Blue Eyes Intelligence Engineering & Sciences Publication Retrieval Number: B11070782S319/19©BEIESP DOI : 10.35940/ijrte.B1107.0782S319 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 are retrieved for further calculations. Users comment score as well as the admin one is recalculated and stored as the new rating for the application. 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 from the store in order to prevent

further user downloads and fake reviews being posted. 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 data which has been given by the user. As in the architecture, it can be logically visualized with the tokenization, 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 initialize the score and the flag as zero. Which means that the initial review based rating is set to zero. This would be modified and changed as per the words that are contained in the database as keys. The flag is that which is almost equal to 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 users rating. This algorithm can be described as in Algorithm

1. SCORE CALCULATION

• input1: user's comment/review given • input2: Single and multikey values • output: Score based on there view • Initialize score=0,flag=0 • Select multikey, singlekey where flag=0 • get the score of single key= entered string • get score of multikey=entered string • score=(singlekey score or multi key score)/2 • return score value 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 ranking fraud detection. The experimental result sshowed the effectiveness of the proposed system, the scalability of detection algorithm as well as some regularity in the ranking fraudactivities

Python:

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design

philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object - oriented approach aim to help programmers write clear, logical code for small- and large-scale projects.

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-oriented way or a functional way.

Flask:

Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

MYSQL:

My SQL is an open-source relational database management system (RDBMS). Its name is a combi nation of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language. A relational database organizes data into one or more data tables in which data types may be related to each other; these relations help structure the data. SQL is a language programmers use to create, modify and extract data from the relational database, as

well as control user access to the database. In addition to relational databases and SQL, an RDBMS like My SQL works with an operating system to implement a relational data base in a computer's storage system, manages users, allows for network access

and facilitates testing database integrity and creation of backups. My SQL is free and open-source software under the terms of the GNU General Public License, and is also available under a variety of proprietary licenses. My SQL was owned and sponsored by the Swedish company My SQL AB, which was bought by Sun Microsystems (now Oracle Corporation). In 2010, when Oracle acquired Sun, Widenius forked the open - source My SQL project to create Maria DB. My SQL has stand - alone clients that allow users to interact directly with a My SQL database using SQL, but more often, My SQL is used with other programs to implement applications that need relational database capability. My SQL is a component of the LAMP web application software stack.

Literature Survey:

The main focus of this project is on emotions analysis and data mining to extract the generated a database. By using this method, we will be able to determine which is true the importance of the apps offered in the Play and App shops r identification fundraising for a single client (i.e., the flexible), Proposed Fraud Measurement Plan Proposed. Testing is available collected to give a as position to each application. Although had identified a variety of sources that were not effectively considering the fact that IP snooping can be done. This IP view allows users to change their IP address and allow them to rate the app more than once. Star ratings are given individually apply application is not sufficient in determining that the app is worth downloading to mobile or not. As he explained [16] that it is wrong to believe in a star rating as can be deceived by engineers themselves. It is considered in reading further reviews there are limits. In general, it is advisable [17] for further testing reliable sources such as selected third-party reviews or to test other developer applications. Collection of specific application databases for a period of time and classify them as good and bad reviews. To use a few words in reviews.

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 value of the applications which are provided nPlay and App stores. Such a proposed system will contain a huge amount of data set that has to be dealt with and using data mining along with visual data will help in carry in gout the system. Information or data mining is the way toward extricating required information from substantial informational collections and changes it into a justifiable arrangement for some time later, essentially utilized for some, 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 data from them. At an exceptionally fundamental dimension, it is discovering extremity of the announcements. Information is gathered from different internet-based life, 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 than once. 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 application recommendation or even focuses on the detection of malwares in the apps before downloading them. Google uses Fair Play 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

Design:

From the Literature Survey and other previous proposals programs designed specifically for this a purpose, the problem in eliminating a fraudulent application is still low function. There are certain functions that involve the use of the web

spam detection rate, online and mobile spam updates the application recommends that you focus and focus on the acquisition of malware in applications before downloading. Google uses Fair Play is a program that detects malwares they only exist in some applications but do not work well enough to do so for the sake of hiding structures. User do not be fooled into downloading the app by its own standards even when it contains certain germs it can infect mobile performance.

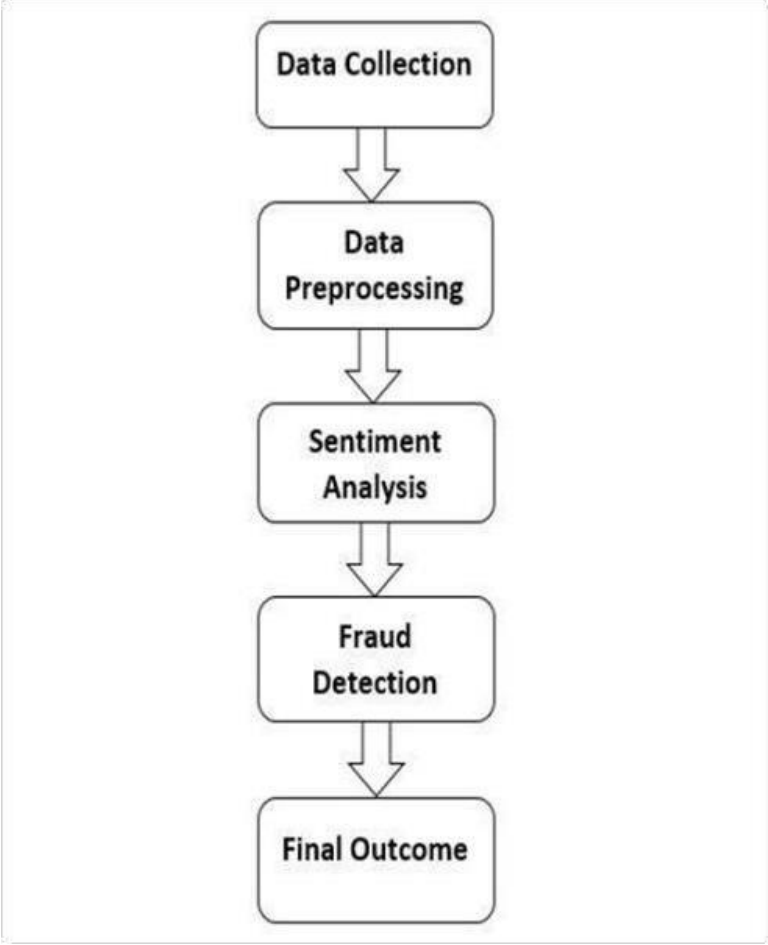
we propose a system that includes identification fraudulent applications use sensitive comments and data mining. We are able to check the sensitivity of the user comment on multiple applications in comparison with fad minimum and user views. looking at these comments, we can distinguish as positive or negative ideas. With a combination of three evidences: based on standard, from the standardized and once based on reviews we are able to find high probability of outcome. Data is extracted and processed by the mining leader sessions. The data was then tested in all three categories evidence and compiled before the final result. Icon important about emotional analysis and data mining in advance continuation over the proposed system and algorithms.

A. Sentiment Analysis: Sentiment Analysis also known as Opinion mining a proper extraction of sensitive and output content emotional data on source material and business assistance in the understanding the social order of their image, object or management while viewing web chats. Emotional analysis is a widely known content a collection device that investigates an incoming message and means that the basic measure is definite, negative or impartiality. Currently, in the emotional analysis is an incredible subject conspiracy and development as it has many advantages applications. Since freely and confidentially the data is accessible over the internet continues to grow, connecting with more people.

B. Data Mining: There is a huge amount of information available at information industry. This an information is no use until it is converted into useful data. Icon it is important to check this large amount of information and incorporate useful data from it. Data extraction is not the main process we have to do; information mines in addition it covers various processes, for example, Data Cleaning, Data

Combination, Data Modification, Data Mining, Pattern Testing and Data Presentation.

Flow Chart:



Architectural Diagram:

Our proposed system as in Fig 1 provides a complete flow of the process that takes place. It starts

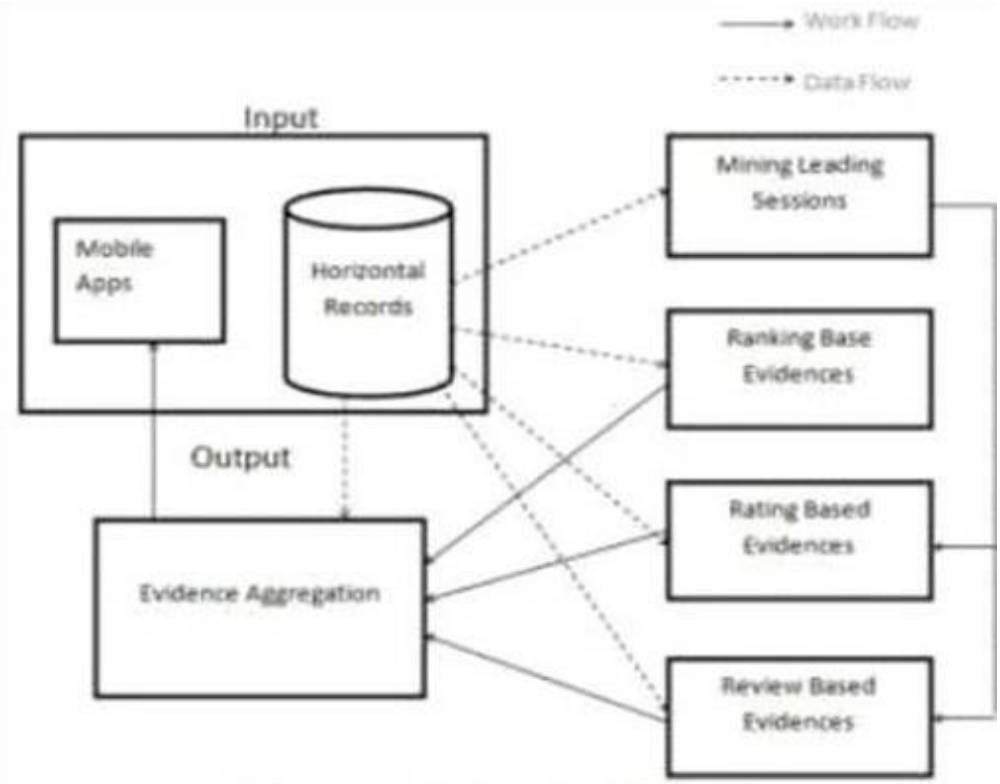


Figure 1. System Architecture

with the output of data which is the history records of applications once store user information. The admin adds an application to database and measurement details. From here, it will mine is a leading session where it is calculated on a basis of the evidence recognized by that particular application. In this case, an advanced mining session algorithm is used that knows identify session and top of an event. After that, evidence rating, status and reviews they are considered individually. A measure of this evidence will be combined with the idea of different time sessions are highly dependent on the main session.

Hardware Design:

The application must provide accurate results.

- Perform the desired function : sorting fraud applications.
- Provide better flexibility and is user friendly.
- User should have to access system to the previous analyzed reports.
- User of the system should have operating systems like Windows 7, Windows8 and Windows10 (32/64 bit).
- The system is implemented using AndroidStudio(JAVA, XML).
- We require minimum 3 GB RAM , 8 GB RAMrecommended, plus 1 GB for the Android Emulator.
- The system should have 1280 x 800 minimum screenresolution.

J48 Algorithm:

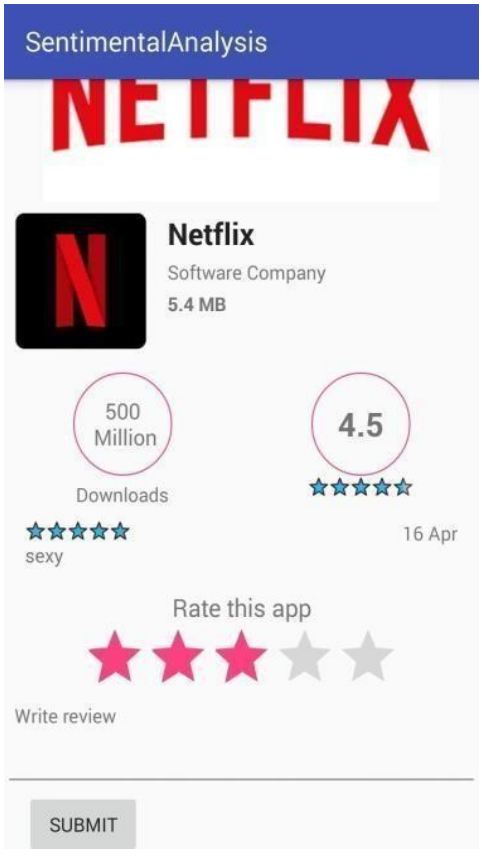
The J48 algorithm is one of the best ways to learn the machine to test data phase and continuously. When used for example purpose, it replaces additional memory and reduces efficiency and the data accuracy in classifying medical data.

It concludes that the leading session has a variety of leading events. Therefore, with the basic of the behavioral analysis of leading events to find evidence of fraud

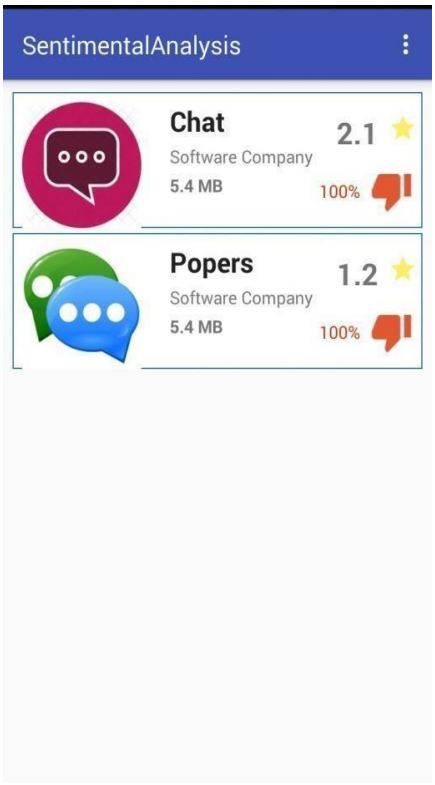
and application - level history from a record, it has been noted that a particular level pattern is always satisfied with the app-level of a behavior at a leading event.

Ranking based Evidence

It concludes that the leading session has a variety of leading events. Therefore, with the basic behavioral analysis of leading events to find evidence of fraud and application - level history records, it has been noted that a particular level pattern is



always satisfied with the app-level behavior at a leading event.


SentimentalAnalysis
Chat
Software Company
5.4 MB
2.1
100%
Popers
Software Company
5.4 MB
1.2
100%

Previously evidence-based evidence is useful for the

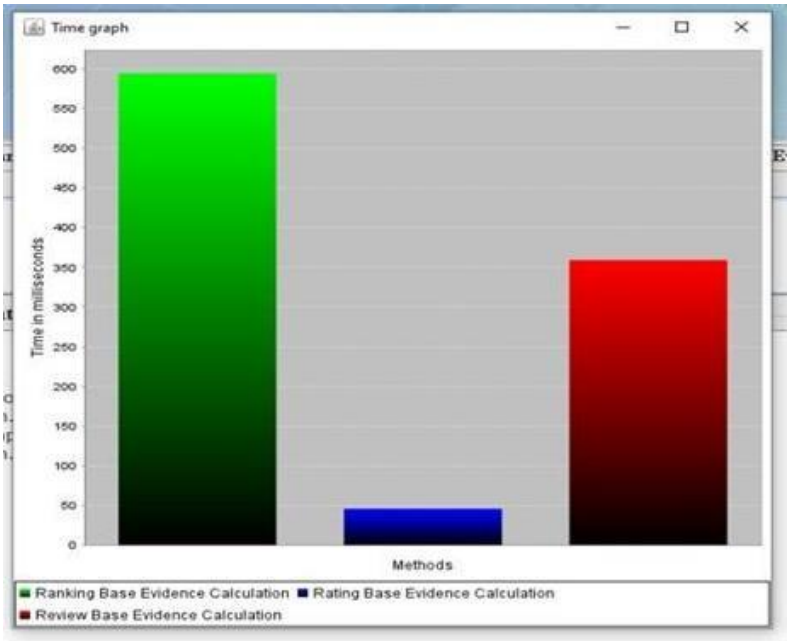
purpose of finding but not enough. To solve the "overtime reduction" problem, fraud proof detection is scheduled based on app history history measurement records. As we know the rating is done after downloading the user, and when the rating is high on the leader board that attracts most users of the mobile app. By default, from the measurements during a moving session reveal a complex pattern that occurs during measurement manipulation. These historical records can be used to develop evidence-based evidence.

Review based Evidences

Evaluation of evidence is a process of data collection and presented in a concise manner. Data may be collected from multiple data sources for the purpose of combining these data sources into the a summary of data analysis.

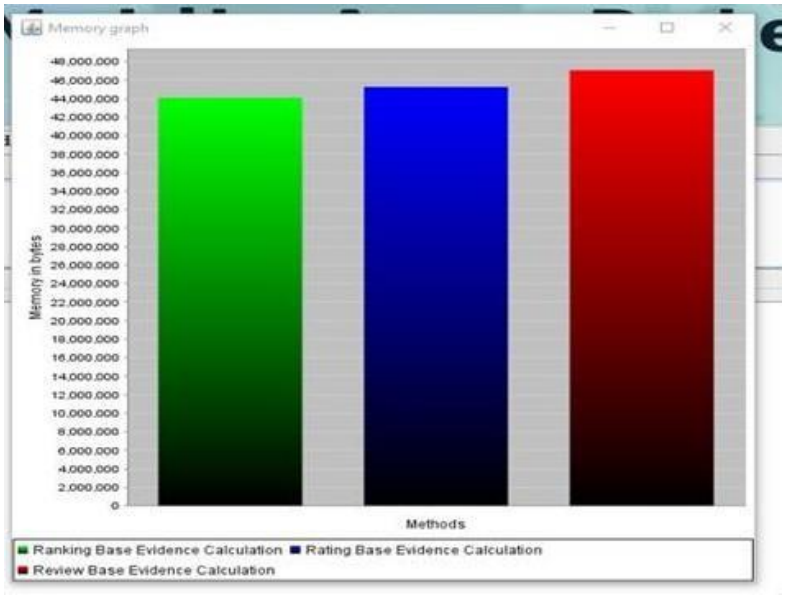
Evidence aggregation

Result Analysis and Comparision:



Ranking, Rating, Review based evidence calculation Time graph

The graph above shows the time taken to calculate the standard, estimate and review based evidence.



Ranking, Rating, Review based evidence calculation Memory graph

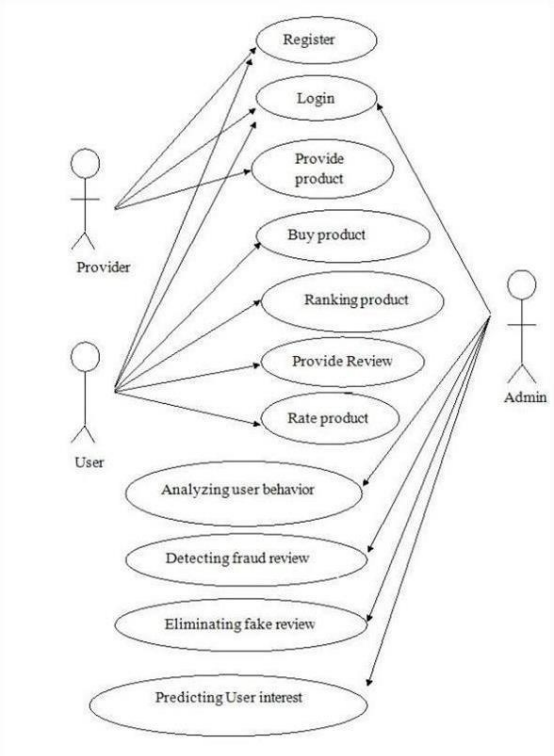
The graph above shows the memory used to calculate the standard, estimation and review based on evidence.

Advantages: The proposed framework is measurable and can be expanded with other evidence a produced by the domain in order to detect fraud by standard.

The test results show the effectiveness of the proposed system, the measurement of the detection algorithm and the specific frequency of fraudulent activities by level.

To the best of our knowledge, it is not a valid indicator of which times are best or which apps actually contain quality fraud.

USE CASE DIAGRAM:



CONCLUSION :

This report presented a presentation on fraudulent applications using the concept of data mining and emotional analysis. It is based on a structural diagram that explained in detail the algorithm and processes used in the project. Data is collected and stored on a re-tested website and defined supporting algorithms. This is a unique way in which evidence is combined and stored for a single outcome. The proposed a framework is measurable and can be expanded to produce more proven domain evidence of fraudulent detection.

8. Conclusion In this project, we have conducted a survey regarding different methodologies used in reviewing the status of application and predicting whether it is fraud or not. Our proposed methodology deals with sentiment analysis which has an advantage over the other methods due to fact that lexicon-based analysis is more accurate and faster than other approaches. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you would like less training data. It is a fast algorithm for classification problems. It is an honest fit real time prediction, multiclass

prediction, recommendation system, text classification and sentiment analysis use cases. Naive Bayes Algorithm are often built using Gaussian, Multi nomial and binomiall distribution. it's very low computational cost.

FUTURE SCOPE:

The Scope of this app is very helpful to the users to use the lot of apps in the daily basis life. And in the many apps we have seen the they are fraud and users cannot find them that they are fraud app or real app so this app will help the users that the app is fraud or not.

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. Fuzail Misarwala, Kausar Mukadam, and Kiran Bhowmick, "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, Mohd FaizalAb RAZAK, "Discovering optimal features using static analysis

and a genetic search based method for Android malware detection" Frontiers of Information Technology and Electronic Engineering, 2018.
5. Javvaji Venkataramaiah, Bommavarapu Sushen, Mano. R,
Dr. Gladispushpa Rathi, "An enhanced mining leading session algorithm for fraud app detection in mobile applications" International Journal of Scientific Research in Engineering., April2017.