SURVEY ON FRAUD APPS DETECTION USING SENTIMENT ANALYSIS

[1]Suhair K M [2]Suhail K M [3]Muhammed Suroor M A [4]Anwar Sadath T M [5]Neha Beegam P E
[1 2 3 4] Student, [5] Assistant professor
[1 2 3 4 5] Computer Science and Engineering, APJ Abdul Kalam Technological University
[1 2 3 4 5] Ilahia College Of Engineering Technology, Muvattupuzha, Kerala, India

Abstract : Smartphones and digital gadgets are beneficial in our daily life. There are such a lot of fraud applications available in the cyber world. In applications fake deportment is the maximum popular. Since there are such a lot of apps available, every one of them could be a scam, it's crucial to identify genuine apps. The proposed system will detect the fraud application based on user evaluations and ratings using the Naive Bayes classifier. The system will avail the utilizer to discover which application is true. The utilizer reviews can be accumulated from the Play Store and relegate reviews into Positive or Negative by utilizing sentiment analysis.

***IndexTerms* - Sentiment analysis, Naive Bayes Classifier, Machine learning, AWS comprehend**

I. INTRODUCTION

With the growth of the app business, corporations and app developers competed to build more features and attract more clients. User reviews, app ratings, and downloads all play a position in determining an app's ranking. There are currently millions of apps available on app stores such as Google Play Store, Apple Store, and others. To transfer an application, every user must go to these play stores. As a result, the review area of every app on the Play Store encourages users to investigate a specific app. Among the application list, the utilizer doesn't belief the appliance if the applications are auxiliary or nugatory. Sometimes the downloaded application won't work or is not subsidiary for the utilizer. This denotes it fraud in the mobile application list.

II. RELATED WORKS

According to Rahul Pradhan, Vedanth Khandelwal, Ankur Chaturvedi, Dilip Kumar Sharma [1] in "Recommendation system using lexicon based sentimental analysis with collaborative filtering" The paper proposes a recommendation system to the customers based on their interest. Sentimental analysis is used to recommend based on the top k rated reviews based on top positive reviews of a geographically wise. But the system cannot predict the correct products all time and cannot recommend products if there is no review.

According to G. Satyanarayana, Dr. Bhuvana J, Balamurugan M [2] in "Sentimental Analysis on voice using AWS Comprehend" By using Amazon Comprehend the system can extract the unstructured data like images, voice etc. Thus, will identify the emotions of the conversation and give the output whether the conversation is Positive, Negative, Neutral, or Mixed. To find sentiment result in this proposed system author is used some AWS services like Identity Access management is used for providing the security to the files.

According to Chenhong Cao Yi Gao,Yang Luo, Mingyuan Xia, Wei Dong, Chun Chen Xue Liu [3] in "AdSherlock: Efficient and Deployable Click Fraud Detection for Mobile Applications" This paper introduces AdSherlock, a client-side click fraud detection technique for mobile apps that is both efficient and deployable. There are two stages to AdSherlock. In the first phase, the offline pattern extractor automatically runs each app and offers a set of traffic patterns for efficient ad request detection. The online fraud detector, as well as the created patterns, are then instrumented into the app and run with it in real-world scenarios.

According to Jaeyoon Kim, Jangwon Seo, Minhyeok Lee, Junhee Seok [4] in "Stock Price Prediction Through the Sentimental Analysis of News Articles" This article performs sentimental analysis by building and analyzing a sentimental dictionary with news articles. Through the sentimental dictionary, we can obtain the positive index of news articles for each date. By analyzing the correlation value between the positive index value and the stock return value, one can confirm the utility and possibility of the sentimental analysis in the stock market.

According to Najma Sultana , Pintu Kumar , Monika Rani Patra , Sourabh Chandra and S.K. Safikul Alam [5] in "Sentiment Analysis for Product Review" The sentiment analysis involves classification of text into three phases - Positive, Negative or Neutral. It analyzes the data and labels the better and worse sentiment as positive and negative respectively. Its merits include Tackles problems on Sentiment analysis, polarity categorization and the demerit includes it does not work well for reviews that purely contain implicit sentiment .

According to Mandava Rama Rao, Nandhini Kannan, CH V S Nihanth [6] in "Detecting Fraud Apps using Sentiment Research" The goal is to employ emotional analysis and data mining to create a system that can detect fraudulent apps before the user downloads them. The system can learn and analyse sentiments and emotions regarding reviews and other writings using sentimental analysis and data mining.

According to K.Manoj, T.S.Sandeep, N Sudhakar Reddy, P.M.D. Alikhan [7] in "Genuine ratings for mobile apps with the support of authenticated users reviews" It primarily provides a perspective of reviews that have an impact on mobile apps or items, as well as detecting fraudulent review lists created by users. There are primarily three evidences to finish up plainly aware of the mobile app scam, namely, Confirmations based on positioning, ratings, and lastly evidence based on a review. The paper pinpoints positioning extortion by analysing dynamic periods in certain driving sessions for mobile apps, as well as discussing how reviews are most beneficial to new clients.

According to Mahmudur Rahman, Mizanur Rahman, Bogdan Carbunar, and Duen Horng Chau [8] in "Search Rank Fraud and Malware Detection in Google Play" The article presents Fair Play, a revolutionary method that detects malware and programmes that have been subjected to search rank fraud by discovering and leveraging traces left behind by fraudsters. In order to identify suspicious apps, Fair Play correlates review actions and uniquely blends discovered review relations with language and behavioural cues collected from Google Play app data. It has a high level of accuracy in detecting both fraudulent and virus Google Play store apps.

III. PROPOSED SYSTEM

The system focuses on classification of application by analyzing user reviews.The sentiment of the user is obtained from the review either as positive negative or neutral. The average rating of the review is obtained from the model and if the average rating of the app has a score above 4 then only the positive reviews are only displayed to the user. If the average rating has a score of less than 4 then all the positive and negative reviews are displayed to the user. The relegation of apps using the Naive Bayes algorithm is the mechanism used in this study. With the increase of the number of online Apps, this project presents a simple and successful framework to distinguish the misrepresentation of Apps. The method calculates a specific app's typicaL rating and compares it to a preset threshold.The ratings which are above three are considered as positive ratings and ratings below three are considered as negative ratings. Determinately, the outcome is within the style of zeros and ones i.e. negative rating offers zero as associate degree output whereas positive rating offers one as associate degree output. After preprocessing of Utilizer Ratings and Reviews Analysis Positive review integrates plus one to a positive score, if negative it will integrate one to a negative score. In this way, it will determine the score of every Utilizer Ratings and Review, determine positive and negative reviews and represent by graph and support positive and negative reviews percentage share whether the app is fraud or not.

Figure 1: Proposed System

The main modules in the proposed system are given below.

3.1 Collecting the app review dataset

The dataset contains a review of applications. There are sundry app datasets available like Google play store dataset from open sources like Kaggle, Google public datasets etc. The dataset contains the reviews for sundry application classes like Gregarious, Games, Inculcation, Finance, News, Aliment.

3.2 Data preprocessing

This step involves the processing of utilizer reviews for the abstraction of dispensable text. Data preprocessing is a process of put together the raw data and making it congruous for a machine learning model. It is the first and key step while engendering a machine learning model.

3.3 Tokenization

Tokenization is the process of breaking a string into smaller parts, such as words, phrases, symbols, other other bits, known as tokens. Tokenization is the process of breaking down large amounts of text into smaller bits. Tokenization divides the raw text into words and sentences, which are referred to as tokens. These tokens aid in the comprehension of elements and the development of a model for NLP. By evaluating the sequence of words, tokenization aids in elucidating the text's construal. Further processes, like as parsing and text mining, use the token list as input.

3.4 Stop Word Removal

In NLP, frivolous words are referred to as stop words. Stop words are the most prevalent terms in any natural language. When reviewing text data and creating NLP models, these stop words will offer no value to the document's interpretation. They can safely be ignored without giving up the meaning of the sentence. If the task is text relegation or sentiment analysis the cessation words should be abstracted as they do not provide any information to our model, but if the task is language translation then stop words are utilizable, as they have to be translated along with other words.

3.5 Stemming

Stemming is the process of lowering a word to its word stem that binds to suffixes and prefixes or the word roots kenned as the lemma. Probing, apperceiving, and retrieving more forms of words return more results. When a form of a word is acknowledged it can make it feasible to return search results that otherwise might have been missed. That additional information retrieved is why stemming is fundamental to probe queries and information retrieval. A stemming algorithm is employed to seek out the base word.

3.6 Algorithm

Sentiment Analysis uses NLP, linguistics, and text analysis to determine the inclination of user opinions. According to ML professionals and developers, SVM and Naive Bayes are the most user-friendly supervised machine learning algorithms. Sentiment analysis gives you an approximate notion of the positive, neutral, and negative attitudes expressed in the texts. The Naive Bayes Classifier Algorithm is employed in this scenario.

3.7 Naive Bayes Classifier

The Naive Bayes classifier is the biggest facile and most customarily used classifier. The Naive Bayes classifier may be a Bayesian learning method and it is utilizable in many practical applications. It is called naive because it includes the simplifying postulation that attribute values are conditionally independent by given the relegation of the instance. The Naive Bayes relegation model computes the posterior probability of a category. It is a supervised learning algorithm, which is put together on the Bayes theorem and utilized for solving relegation quandaries. Naive Bayes Classifier is one of the most facile and most efficacious Relegation algorithms which avail in building expeditious machine learning models that can make expeditious prognostications. Bayes' theorem is additionally kenned as Bayes' law, or Bayes' Rule which is utilized to decide the probability of a theorem with precedent cognizance. It depends on the conditional probability.

The formula for Bayes' theorem is given as follows:

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)} \qquad (1)$$

Where,

P(X|Y) is known as Posterior probability: Probability of hypothesis X on the observed event Y.
P(Y|X) is known as Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(X) is known as Prior Probability: Probability of hypothesis before observing the evidence.
P(Y) is known as Marginal Probability: Probability of Evidence.

Naive Bayes Classifier is a kind of classifier that works on Bayes theorem. Forcast of membership probabilities is made for every class such as the probability of data points associated with a particular class. The class having maximum probability is appraised as the fittest class. This is also termed as Maximum A Posteriori (MAP).
The MAP for a hypothesis is:

- ➢ $MAP(H)=\max P((H|E))$
- ➢ $MAP(H)=\max P((H|E) * (P(H))| P(E))$
- ➢ $MAP(H)=\max (P(E|H)*P(H))$

$P\ (E\ )$ is evidence probability, and it is used to normalize the result. The result will not be affected by removing $(E\ )$. Apart from considering the freedom of every feature, Naive Bayes also concludes that they contribute equally.
We can recast the Bayes Theorem for our example as:

$$\mathrm{P(y \mid X)} = \frac{[\mathrm{P(X \mid y)\ P(y)\ P(X)}]}{\mathrm{P(X)}} \qquad (2)$$

Here, y stands for the class variable. X stands for the features.

$$X = (x_1, x_2, x_3, \ldots., x_n)$$

Now, we have to replace X and fill out the chain rule to get the following:

$$\mathrm{P(\,y \mid x_1,\ldots,x_n\,)} = \frac{[\,\mathrm{P(x1 \mid y)\ P\ (x2 \mid y)\ \ldots.P(xn \mid y)\ P(y)}\,]}{[\,\mathrm{P(x1)\ P(x2)\ldots.P(xn)}\,]} \qquad (3)$$

One can get the values for each by using the dataset and putting their values in the equation. The denominator will remain the same for every entry in the dataset to remove it and inject proportionality.

$$\mathrm{P(y \mid x_1,\ldots,x_n)} \propto \mathrm{P(y)\pi^n{}_{i=1}P(xi \mid y)} \qquad (4)$$

Now, it is essential to create a classifier model. For this, first, find the probability of a given set of all possible inputs for all possible values of the class variable y and pick up the output with the highest probability. This can be expressed mathematically as:

$$\mathrm{y\ = argmax_y\ P(y)\pi^n{}_{i=1}P(xi \mid y)} \qquad (5)$$

P(y) is called class probability and P(xi | y) is called conditional probability. The result of the prediction would be the class that has the highest posterior probability

IV. CONCLUSION

As technology evolves, so does people's thinking, and certain applications may be fraudulent. With regards to offering facilities to clients, the protection of the client is the main thought. Such an application could be virus or information burglary, and there should be some sort of security set up profiting hand to manage the utilizer about any early application. This application not only suggests but additionally provides security to the utilizer in a better way.In the future, it's orchestrated to review more practical fraud evidence and analyze the latent relationship among rating, review and rankings. Besides, it will be elongated to ranking fraud detection approaches with alternative mobile App cognate accommodations, like mobile Apps recommendation, for enhancing utilizer experience.

V. ACKNOWLEDGEMENT

Aside from the endeavors of us, the accomplishment of this present venture's primer report relies generally upon the consolation and rules of numerous others. We might want to accept this open door to thank each and every individual who added to the fruitful finishing of this project. We might want to show our sincere appreciation towards Prof. Dr. ABDUL GAFUR M, Principal, Ilahia College of Engineering and Technology for allowing us to deal with this project. Likewise, we want to show our most noteworthy appreciation towards our head of the Department of Computer Science & Engineering Dr. Arun E and project guide Ms. Neha Beegam P E for their significant exhortation and direction. At last, we offer our thanks and gratitude to every one of our educators and other employees of the Department of Computer Science & Engineering, for their genuine and agreeable participation in finishing this task.

REFERENCES

[1] Rahul Pradhan, Vedanth Khandelwal, Ankur Chaturvedi, Dilip Kumar Sharma, “Recommendation System using Lexicon Based Sentimental Analysis with collaborative filtering”, International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), 10.1109/parc49193.2020.236571.

[2] G. Satyanarayana, Dr. Bhuvana J, Balamurugan M, “Sentimental Analysis on voice using AWS Comprehend”, International Conference on Computer Communication and Informatics (ICCCI 2020),10.1109/iccci48352.2020.9104105.

[3] Chenhong Cao Yi Gao, Yang Luo, Mingyuan Xia, Wei Dong, Chun Chen Xue Liu, “AdSherlock: Efficient and Deployable Click Fraud Detection for Mobile Applications”, IEEE Transactions on Mobile Computing, 10.1109/tmc.2020.2966991.

[4] Jaeyoon Kim, Jangwon Seo, Minhyeok Lee, Junhee Seok, “Stock Price Prediction Through the Sentimental Analysis of News Articles”, Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), 10.1109/icufn.2019.8806182.

[5] Najma Sultana , Pintu Kumar , Monika Rani Patra , Sourabh Chandra and S.K. Safikul Alam, “Sentiment Analysis for Product Review”, International Journal of Soft Computing, 10.21917/ijsc.2019.0266.

[6] Mandava Rama Rao, Nandhini Kannan, CH V S Nihanth, “Detecting Fraud Apps Using Sentiment Research”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S3

[7] K.Manoj, T.S.Sandeep, N Sudhakar Reddy, P.M.D.Alikhan, “Genuine ratings for mobile apps with the support of authenticated users reviews”, Second International Conference on Green Computing and Internet of Things (ICGCIoT), 978-1-5386- 5657-0/18.

[8] Mahmudur Rahman, Mizanur Rahman, Bogdan Carbunar, Duen Horng Chau, “Search Rank Fraud and Malware Detection in Google Play”, Transactions on Knowledge and Data Engineering PP(99):1-1, 10.1109/TKDE.2017.2667658.

[9] Shraddha Jundhare, Padmaja Gajare, Priyanka Gadekar, Archana Aher, Shalini Wankhade, ”Fraud Application Detection Using Summary Risk Score”, International Conference on Inventive Systems and Control (ICISC),10.1109/ICISC.2017.8068694.

[10] Varsha A. Patil, and Nitin N. Patil, “Mobile Apps Opinion Analysis UsingEmoticon”, International Conference on Global Trends in Signal Processing, Information Computing and Communication, 978-1-5090-0467-6/16.

[11] Hengshu Zhu, Yong Ge, Enhong Chen (2015) , “Discovery of Ranking Fraud for Mobile Apps”, IEEE Transactions on Knowledge and Data Engineering, 10.1109/TKDE.2014.2320733.

[12] Xujuan Zhou, Xiaohui Tao, Jianming Yong, Zhenyu Yang, “Sentiment Analysis on Tweets for Social Events”, IEEE 17th International Conference on Computer Supported Cooperative Work in Design, 978-1-4673-6085-2/13.