A Novel Approach for clone app detection using VADER's Algorithm

Sonali Antad and Shreyas Khamkar, Ayush Khambayate, Om Khandare, Atharva Khaire, Prasad Khambadkar, Deep Khanchandni

Department of Engineering, Sciences, and Humanities (DESH) Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

sonali.antad@vit.edu, ayush.khambayate22@vit.edu, shreyas.khamkar22@vit.edu, om.khandare22@vit.edu, atharav.khaire22@vit.edu, prasad.khambadkar22@vit.edu, deep.khanchandani22@vit.edu

Abstract. Software that imitates the capabilities of legitimate, legally responsible, and authentic applications is knownas a fraudulent web application. It's critical to keep track of which mobile applications are secure and which aren't as the number of them in our daily lives increases. One cannot judge the truth, and the only basis for judging each application is the opposing viewpoints that are stated for each application. Once the false program has been installed, the perpetrators carry out retaliatory acts such as aggressive ad display to recoup revenue, intercepting sensitive data from your system, polluting the impact device, and so forth. Users frequently cannot tell the difference between legitimate and fraudulent applications By developing a place where people can ask questions before installing the program, It is suggested to employ sentiment analysis (VADERS), which is a revolutionary method for detecting fraudulent apps.. The outcome is determined by the ratings and comments provided by users who have already used the application. As a result, we will use sentiment analysis to examine the viewpoints once more. Sentiment analysis will be carried out using the VADERS approach, which analyses text.

Keywords: Sentiment analysis, Positive – Negative Ratings, App Security, Machine learning Models, Behavioral Feature Extraction.

1 Introduction

The use of mobile devices has increased along with technology. The creation of different kinds of mobile applications for platforms like Android and iOS has increased significantly.

In Covid's circumstances, growth is basically at an all-time high. The world of the business intelligence industry is facing a sizable problem as a result of this technology's daily rapid growth in terms of its uses, modifications, and development. The result is increased market competition. Due to the intense rivalry this company and many application developers face one another, they put in a tremendous lot of effort to draw in new clients, keep those clients once they have been recruited, and support their continued growth.

Customers' rankings, evaluations, and opinions regarding the particular program they download are of utmost importance. This might be a technique for both new and experienced developers to identify their areas for improvement while creating a new product with the needs of the user in mind.

In order to accurately identify the ranking scam, we first recommend mining the busy times by abusing mining the top session algorithmic program. Additionally, we typically look at three other sorts of evidence when analyzing historical records: ranking-based, rating-based, and once again, view-based data. Last but not least, we typically respect the anticipated system with knowledge of real-world apps obtained from the Google Play Store over a lengthy period oftime. We will show the significance of the cognitive algorithmic program in the tests along with some consistency of the positioning misstatement exercises, and we will generally confirm the validity of the precedent framework. The majority of fraud acknowledgment frameworks categorize views and assessments of the applications into two groups, i.e. Extremely good, good, neutral, bad, and very bad.However, due to mixed second opinions, several ratings and second opinions are not grouped into significant groups.

2 Literature Review

The software development system seeks to recognize clone apps before users download them by utilizing sentimental analysis and data mining. [3].

Sentiment analysis is a method in which we can detect the mood or emotional state of the person who is writing the reviews whether the person is happy or sad while writing thereview. The article discusses the issue of ranking fraud

in themobile app market, where users may download apps based onmisleading rankings and end up with useless or non-functioning apps [2].

The authors propose a ranking fraud detection system that uses a mining leading session algorithm to detect active

periods and investigates three types of evidence - ranking, rating, and review-based - to integrate and detect fraud [4].

The method is tested using actual Google App Store data, and the findings demonstrate the efficiency and scalability of the suggested algorithm in identifying ranking fraud [14].

The article discusses positioning misrepresentation in the mobile app market, where developers use shady means to increase their app's ranking and popularity [8].

The authors propose a positioning fraud detection system using data mining and sentiment analysis techniques to analyze app data and determine if fraudulent activities are present [2].

Sentiment analysis establishes a piece of literature's positive, negative, or neutral tone, whereas data mining analyses data from several angles to extract usable information. By combining these techniques, the proposed system can detect and prevent positioning misrepresentation in the mobile app market.

3 Methodology/Experimental

1.] A large collection of data from the Google Play Store has been extracted. The review is copied manually. User feedback is gathered for 4 different applications kinds.:

1. Social
2. Shopping
3. Credit card transition fraud
4. Application fraud

2.] In the project technologies like “Machine Learning” and “Sentimental Analysis”. The software used in this projectis SQL, ADVANCED HTML 5.

The proposed approach for clone app detection using sentiment analysis involves the following steps:

1. Collecting app reviews: Utilizing the "google_play_scraper" library, app reviews are gathered. The app "com.edurev.class1" is the best option for investigation. The "reviews_all" function was used by all views, which were then arranged in ascending order. look date.

2. Data preprocessing: The JSON format of the views data was transformed into a pandas data frame using the "pd.json normalize" function. Additionally, the content column was converted into a string type for easier sentiment analysis.

3. Sentiment analysis: The VADERS model was employed for sentiment analysis on reviews. The sentiment analysis was applied to the content column of the data frame using the "apply" function. From the analysis results, we extracted the sentiment label and score, which were then appended as separate columns to the data frame.

4. Data visualization: The sentiment analysis results were visualized using the "plotly.express" library to generate a histogram. The y-axis represented the percentage of views in each sentiment category (positive, negative, or neutral), while the x-axis denoted the sentiment category itself. To modify the y-axis label to "percentage", the "update_layout" function is used.

Libraries used: The following libraries were used for Implementation:

PLOTLY.EXPRESS:
Plotly. express is a Python library used for creating interactive data visualizations. It provides a high-level interface for creating charts and graphs with minimal coding.It supports various chart types including scatter plots, line charts, bar charts, and histograms. It is widely used in data science and machine learning projects for data exploration and presentation.
Vader Sentiment Analyzer is a natural language processing tool that is used to analyze the sentiment of a piece of text. It is an open-source tool developed by researchers at the Georgia Institute of Technology, and it uses a lexicon of words and their scores to determine the sentiment of a text. Vader Sentiment Analyzer is unique because it can analyze sentiment in a way that takes into account the intensity of emotions and the context in which the words are used. This makes it particularly useful for analyzing social media data, where context is often key in understanding the sentiment of a post or comment.
The lexicon used by Vader Sentiment Analyzer consists of words that are rated on a scale from -4 to +4, with -4 being extremely negative and +4 being extremely positive. The lexicon also includes words that are considered neutral, such as "the" and "and." Vader Sentiment Analyzer takes these ratings into account when analyzing a piece of text.
Vader Sentiment Analyzer also takes into account the intensity of emotions in a piece of text. For example, the word"hate" has a much stronger negative connotation than the word "dislike." Vader Sentiment Analyzer takes these differences in intensity into account when analyzing sentiment.
Another key feature of Vader Sentiment Analyzer is its abilityto handle negations and punctuation. For example, the sentence "I do not like this product" would be analyzed as

negative, even though the word "like" is usually associated with a positive sentiment. This is because Vader Sentiment Analyzer recognizes the negation in the sentence.

Vader Sentiment Analyzer is widely used in social media analysis, marketing research, and customer feedback analysis. It can be used to analyze customer reviews of products, monitor social media sentiment about a brand, or analyze the sentiment of political speeches.

One of the main advantages of Vader Sentiment Analyzer is that it is open-source and freely available. This makes it accessible to researchers and analysts who may not have access to expensive sentiment analysis tools. Additionally, Vader Sentiment Analyzer has been shown to be highly accurate in a number of studies.

However, there are some limitations to Vader Sentiment Analyzer. Like all sentiment analysis tools, it may struggle with sarcasm or irony, which can be difficult to detect in text. Additionally, Vader Sentiment Analyzer may not work well with languages other than English, as the lexicon is based on English words.

In conclusion, Vader Sentiment Analyzer is a powerful tool for analyzing sentiment in text. Its ability to take into account the intensity of emotions and the context in which words are used makes it particularly useful in social media analysis and customer feedback analysis. While it has some limitations, it is a highly accurate and accessible tool that has been widely adopted in research and business communities.

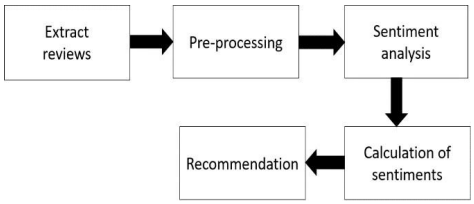

Fig. 1. Flow Diagram of System.

MATH

Average of fraud ness = Sum of ratings received from user /number of voters.

By dividing the total number of votes by the sum of user evaluations, the average fraud score can be determined.

4. Results and Discussions

The bogus web software's sign-in page, or home page, looks like the one below. The option to sign up is also available on the home page. On this website, you may see details about our software, contact information, our company, our services, and our blog.

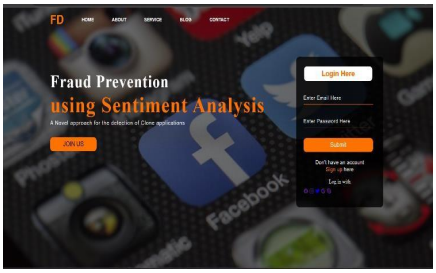

Fig. 2. Login page of the System

For users who have never used this software before, there is an option to sign up after the login page. After selecting this option, the sign-up page opens right away, where the user must fill out their information before clicking the sign-in button.

The project's findings on fraud app detection using sentiment analysis showed how well the method worked to spot possible fraud in mobile applications. The fraud detection algorithm was successful in detecting suspicious reviews and fraudulent applications, while the sentiment analysis model was able to accurately categorize user evaluations as positive, negative, or neutral. The web app provided a user-friendly interface for users to easily check theauthenticity of a mobile application, reducing the risk of beingscammed. The system was able to detect fraudulent applications in real time.

To use the web app first we have to sign up and then log in after which you will be directed to the input page.

Input: App Id
Example: Safe app

After opening the webpage you can see an option of “Input the ID of the app” There you have to fill in the id of the app that you want to see is safe or not. Now, where can you find the ID of an app it’s very simple just go on Google - type the app name – go to the page of google play store where you can see the app, and in the URL of that webpage you will see an ID of an app you just have to copy and paste it in the text space provided and submit the form will now load and it will take time to load as it depends on the number of reviews the app contain if the app contains more number of reviews then it will take more time as the input will load the dataset of reviews in the machine learning model and then sentiment analysis will start and then the output will be displayed

Output:

On the output page the result of the sentiment analysis is displayed and additionally with some instructions and theresult that is the app safe or not for use and it is the recommendation of the sentiment analysis additionally with a graph is also displayed in the next tab of the browser which describes the number of positive and negative reviews.

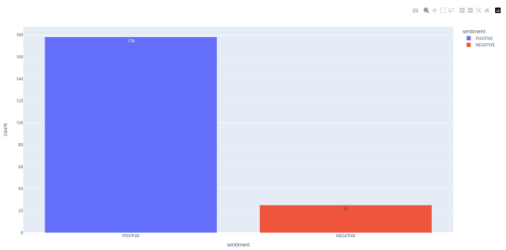

Fig. 3. Graph of +ve ratings

The graph describes the positive and negative number of thereviews of the app so we can get an idea that if the app is reallygood, Positive is represented by the blue bar, and negative bythe red bar reviews which were decided by the sentiment analysis algorithm.

THE RATING and THE GRAPH is sucessfully displayed

Results : The app is Safe to download .

Fig. 4. Final result.

Some examples of safe apps are given as

follows:1] Class 1 CBSE App + Worksheets
2] Vocabulary - Learn fresh
words 3] Deep stash: Smarter
Every Day!4] Medium
Example: Fake app
In this app negative ratings is average half of the positiveratings as shown in fig.6. that's why this app is not safe.

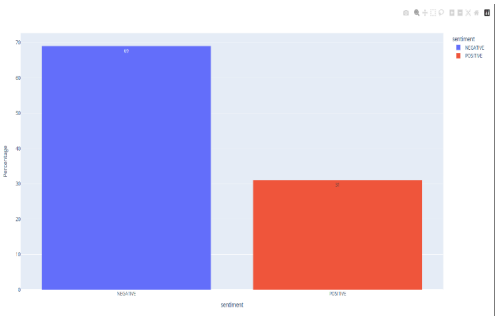

Fig. 5. Graph of –ve ratings

THE RATING and THE GRAPH is sucessfully displayed

Results : The app is not Safe to download, Look far an alternative .

Fig. 6. Final result.

Fig.7. shows whether the final result of this app is fraud or not. This app result shows the app is not safe to download and also shows that you can look for an alternative app.

The efficiency of the VADERs algorithm is more than most of the algorithms which are existing algorithms :

[1] Novel Approach for Fraud App Detection Using SentimentAnalysis

Methodology :
Natural Language Processing, Sentiment Analysis, and VADERS Sentiment Analysis, which has a maximum efficiency of 96%

[2] Information Extraction for Mobile Application User Review.

Methodology :
SVM Logistic Regression Non-Negative Matrix, Sentiment Analysis, Content Classification Topic Modelling, Filtering Latent Dirichlet distribution has an efficiency of 80.50%

[3] An Implementation to Detect Fraud App Using FuzzyLogic

Methodology :
Ontology, fuzzy logic algorithm, and fuzzy logic tokenizationhave an efficiency of 83.75%

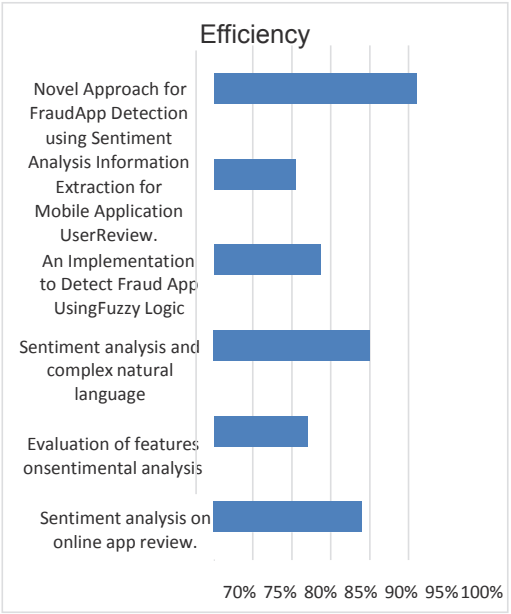

Fig.7. Comparison Efficiency Graph

[3] An Implementation to Detect Fraud App Using FuzzyLogic
Methodology :
Ontology, fuzzy logic algorithm, and fuzzy logic tokenizationhave an efficiency of 83.75%

[4] Sentiment Analysis and complex natural languageMethodology :
Naive Bayes is an NLP Sentiment algorithm. has an efficiency of 90%

[5] Evaluation of features on sentimental analysisMethodology :
Porter's algorithm for stemming and supporting vector machines has an efficiency of 82%

[6] Sentiment analysis on online app review.Methodology :
NLP for machine learning-, and K for cluster have an efficiency of 89%

Acknowledgment

We sincerely thank everyone who helped finish this research work and for their contributions. The assistance and cooperation from experts in the subject have proven invaluable. This work has been significantly influenced by the dedicated efforts, thoughtful observations, and knowledge that researchers, academics, and professionals have contributed. We thank the reviewers for their insightful criticism and useful recommendations, which have greatly improved the caliber and rigor of this study. Additionally, we would like to thank the institutions and groups who donated funds and other resources to make this study possible. Their assistance has been crucial in making this research paper possible to finish.

References

[1] "Mob Safe: Forensic Analysis for Android Applications and Detection of Fraud Apps," vol. 4, no. 10, pp. 3779–3782,2015; p. Rohini, k. Pallavi, j. Pournima, k. Kucheta, and p. P. Agarkar
[2] Prof. Omkar Dudhbure, Dewanand kapgate, Nidhi Nikhar,Ashwini Tichkule, "Revelation of fraud applications usingsentiment analysis app reviews," vol. 4, no. 5, 2019.
[3] Salini, dhevadharshini, and Malath, "Detecting fraud appusing sentiment analysis," vol. 10, no. 7, July 2021. https://ijarcce.com/papers/detecting-fraud-app-using-sentiment-analysis/
[4] "detecting fraud applications using sentiment research," Mandava rama rao, Nandini Kannan, and ch v s nihanth, ISSN: 2277-3878, volume 8, issue 2s3, July 2019. https://www.ijrte.org/wp-content/uploads/papers/v8i2S3/B11070782S319.pdf

[5] Valence arousal similarity-based recommendation services, IEEE international conference on Circuit, Power andComputing technologies, ICCPCT 2015. Dr. R. Subhashini and akila g.

[6] Mobile application for malware detection, Pranjali Deshmukh and Pankaj Agarkar, International Research Journal of Engineering and Technology (IRJET), volume: 02issue: 02 | May 2015. http://irjet.net/archives/V2/i2/Irjet-v2i2161.pdf

[7] 'Emerging trends in engineering & technology' 9. Manoharbhai Patel organized the event in Shahapur, Bhandara, institute of Engineering and Technology 2019's vol. 4, no. 5 of the international journal of Innovations inEngineering and Science.

[8] Optimal aggregation method for fraud detection andprevention in mobile apps, international journal of advanced research in computer science and software engineering, no. 8, march 2016. Pratik phapale, pratik sapkal, dr. Swati jaiswal, laxman kuhile, and vivek pingale. https://www.semanticscholar.org/paper/Fraud-Detection-%26-Prevention-of-Mobile-Apps-using-Pingale-Kuhile/33960b8f62de8811d349d8bdbe2c8c36648b5cd7

[9] Valence arousal similarity-based recommendation services, IEEE international conference on Circuit, Power andComputing technologies, ICCPCT 2015. Dr. R. Subhashini and akila g. https://www.semanticscholar.org/paper/Valence-arousal-similarity-based-recommendation-Subhashini-Akila/c362fd9e96bac69d73deee614e59ff31730e151b

[10] Android malware detection using parallel machine learning classifiers, 8th international conference on next generation mobile apps, services and technology, September2014. Suleiman y. Yerima, sakir sezer, and igor muttik.

[11] M. M. Mhatre, m. S. Mhatre, m. D. Dhemre, and p. S. T. V, "Detection of ranking fraud in mobile applications," pp. 2187–2191,2018. https://www.sciencedirect.com/science/article/abs/pii/S1549963409001154

[12] P. Rohini, k. Pallavi, j. Pournima, k. Kucheta, and p. P.Agarkar, "Mob safe : forensic analysis for Android applications and detection of fraud apps using cloud stack anddata mining," vol. 4, no. 10, pp. 3779– 3782, 2015. https://issuu.com/irjet/docs/irjet-v6i2395/3

[13] Gladence, L. Mary, M. Karthi, and V. Maria Anu. "Astatistical comparison of logistic regression and differentbayes classification methods for machine learning." arpnJournal of Engineering and Applied Sciences 10, No. 14 (2015): 5947-5953. https://www.researchgate.net/publication/282921131_A_statistical_comparison_of_logistic_regression_and_different_bayes_classification_methods_for_machine_learning

[14] Neha M. Puram and Kavita R. Singh, "Semantic analysisof app review for fraud detection using fuzzy logic",International Journal of Computer & Mathematical Sciences, vol. 7, January 2018. https://www.semanticscholar.org/paper/Semantic-Analysis-of-App-Review-for-Fraud-Detection-Puram-Singh/ed73761ad92b9c9914c8c5c780dc1b57ab6f49e8