Semantic Analysis of App Review for Fraud Detection using Fuzzy Logic

Neha M. Puram
Yeshwantrao Chavan College of Engineering, Nagpur
Kavita R. Singh
Yeshwantrao Chavan College of Engineering, Nagpur

ABSTRACT—Now-a-days, use of apps has increased with the increasing craze towards mobiles. For all types of mobile application, users are preferring smartphones. Generally, depending on how many users already have downloaded that application, what are the ratings and reviews, what are the comments, etc., users download mobile applications. In the mobile app market, fraud ranking points to false activity that has a reason to push up the mobile apps in the popularity list. Certainly, it turns more periodic for app developers to use fake mechanism. Here, the paper proposes semantic analysis of app review for fraud detection in mobile apps. Firstly we propose to correctly detect the misrepresentation by excavating the active periods, also called as leading sessions, of the mobile apps. Furthermore, we will inspect two types of evidences, namely, ranking-based evidences, review-based evidences and use natural language processing (NLP) to get action words. Next, convert review to ratings and finally perform pattern analysis on session with app data gathered from the app store. So, the paper proposes approach to validate its effectiveness, and also show the scalability of the detection algorithm.

KEYWORDS—App Review, Ranking Based Evidences, Review Based Evidences, Fraud Detection, Session Based Analysis, Fuzzy Logic.

INTRODUCTION
With the fast progress of smartphones, mobile applications of various class, such as social, financial, game, lifestyle, etc., are available for download from various application markets, like Android’s Google Play Store. Android applications usually supply precised data which include its name, description, category that they belong to, and the rating from users, etc. However, the total number of apps available at Google Play app store are 2,800,000 and in Apple app store is 2,200,000 [1]. The app store daily launch their apps on leaderboard. Leaderboard is one of the way through which apps can be promoted. Higher the rank of app in leaderboard, higher is the number of download. High number of downloads results in millions of dollar in revenue. However, to increase the revenue, most app developers try false means to boost their apps in the leaderboard. Although, some app developers search for different ideas to promote their apps and have higher rank in the leaderboard.

Moreover, the need to develop review fraud detection system for mobile app using fuzzy logic is to increase its accuracy and proper differentiation between the reviews. As, textual comments are allowed by most of the app stores, these comments called as reviews will be used for identifying if the app is genuine or fake. The use of sessions will be proposed to identify the review in that particular session. The ratings will then undergo through mean rating analysis. Fuzzy logic will help to classify the review text as very bad, bad, neutral, good and very good. Also, NLP will help to extract the required action words from the text using POS tagging and chunking. This extracted result will undergo through analysis and comparison, and finally the output will be analyzed.

Few of the main objectives that have been highlighted in this paper are as mentioned below:

- To improve the differentiation between the reviews.
- To enhance the accuracy of classification in a system.

The paper constitutes four sections. Section I has introduced about fraud app and Section II is dedicated to
study on various types techniques used to detect fraud mobile apps. At the end, Section III provides proposed methodology and Section IV provides conclusion.

RELATED WORK

In this section, the paper provides detail survey of the methods used by many researchers for identifying fraud mobile apps. Many deceptive behaviors occur in well-known Android app market i.e., Google Play Store. So, to detect malware, previously the work only focused on permission analysis and app executable. Although, Mahmudur Rahman et. al., introduced FairPlay, a system which detected and leveraged trace leave behind by fraudulent to identify the malware and also the apps subjected to search rank fraud [2]. FairPlay correlated review activity and identified their relation with linguistic and behavioral signals that have been gathered from Google Play app data to detect the suspicious apps. Day by day use of mobile has increased. Also to access all types of mobile application, the mobile users prefer to use smartphones. Users generally download mobile applications depending on how many users already have downloaded that application?, what are its ratings and reviews?, what are the comments? etc. Fraud ranking in the app market indicates false or wrong deeds that might have reason to push up apps on the popularity list. Most app developers use fraud means to increase their app's sales by notifying false ratings of the apps, and carrying out ranking fraud. As well as, Varsha A. Patil et. al., presented work on opinion research on emoticons which is a string of symbols representing different faces in text-based communication [3]. Also an optimization-based aggregation method has been demonstrated, and opinion analysis has been used to find how much a review is positive or negative. The review score has been used to raise the rating score of the user and the emoticons in the reviews or comments. Whereas, Alexis Silva et. al., in [4] introduced a system named BehaviorDroid to track the general properties of apps during run-time. These general properties were currently specified with the help of automata that helped to differentiate between the wanted and undesirable interactions between the resources of phone and app.

However, in paper [5] Josh Jia Ching Ying et. al., proposed an extremely effective fraud phone call detection approach called parallelized graph-mining, namely PFraudDetector. It automatically labeled deceptive phone numbers by the tag named “fraud”, so as to differentiate the fake phone call numbers from the genuine ones. It also used Hyperlink-Induced Topic Search (HITS) algorithm and a novel aggregation approach. In paper [6], Navdeep Singh et. al., presented an optimization-based aggregation approach to incorporate the authentication so as to analyze the capacity of main particular time span from mobile application.

Although, Hengshu Zhu et. al., presented a comprehensive approach for ranking fraud and also for detecting ranking fraud in the mobile apps. Firstly, an active period has been mined accurately to locate the ranking fraud [7] [14]. Furthermore, ranking-based evidences, rating-based evidences and review-based evidences were investigated. These three proofs were investigated by representing the apps ranking, rating and review behaviors with the help of statistical hypotheses tests. Also, an optimization-based aggregation method has been proposed that will incorporate entire evidence of fraud detection. The recommended approach was estimated with the real-world app data which was collected from the iOS app store. Spam web pages posed great challenge to the development of search engine. The content spam was commonly used. Along with the development of Internet technologies, the content spam was difficult to detect. For this reason, Jing Wan et. al., [8] has been proposed a detection method for the web page using content spam technique that primarily relies on the statistical features. A spam webpage detection method based on topic and semantics was proposed, with the use of two categories of features, namely, semantics and statistics.

To model the information of mobile apps against mobile app services, Hengshu Zhu et. al., presented a sequential approach which is based on hidden markov model (HMM) in [9]. Specifically, first popularity based HMM (PHMM) has been presented to model the sequences of the heterogeneous popularity observations of mobile apps. Also, a method named bipartite-based method has been introduced to precluster the popularity observations. Thus it effectively helped to learn the parameters and initial values of PHMM.

Author Siqi Ma et. al., of paper [10] has been proposed an active and semi-supervised technique to detect the malwares. The approach made use of both known harmless and mischievous apps to predict other malignant apps. The proposed approach was also able to select a good set of apps for experts to label as malicious or
harmless to form a set of labeled training data. Furthermore, the approach used both labeled data and unlabeled data, which is a semi-supervised approach. Mayank Taneja et. al., introduced a novel approach for prediction of click fraud in mobile advertising which comprised of feature selection using Recursive Feature Elimination (RFE) and classifies using Hellinger Distance Decision Tree (HDDT) in [11]. RFE was chosen for feature selection as it provided better results as compared to wrapper approach when evaluated using different classifiers. HDDT has been also selected as classifier so as to deal with class imbalance issue present in the data set. The efficiency of proposed framework was investigated on the data set provided by Buzzcity [18] and compared with J48, Rep Tree, logitboost, and random forest.

Kaiyu Wang et. al., presented detection approach in [12] from user’s viewpoint, which combined the features of social network from publishers with textual features of microblogs itself together, to comprise the vector of feature. Then, by providing the vector of feature to the SVM for training data, spam microblogs were classified. The dataset of Sina Weibo has been used for the experiment, which is the most renowned Chinese microblogs, to verify the effectiveness of proposed system. The proposed approach provided 13% and 29% increased accuracy. Hengshu Zhu et. al., in [13] presented a system to enrich the contextual info of mobile apps by employing more web knowledge from the web search engine. Some of the contextual features for mobile apps were also extracted from the context-rich device logs of mobile users. Conclusively, all information was then combined to maximum entropy model for training a mobile app classifier. Large experimentations were carried out on 443 mobile user’s device logs to present the effectiveness and efficiency of the proposed approach. The outcomes of the experimentation presented that the proposed method outperforms two state-of-the-art benchmark methods with a significant margin.

Enhong Chen et. al., illustrated an approach in [15] about extraction of personal context-aware preferences from the context-rich device logs to build novel personalized context-aware recommender system. The proposed work initiated with learning some common context-aware preferences from the context logs of many users. Formerly, the preference of each user has been represented as a distribution of these common context-aware preferences. Precisely, two approaches for mining common context-aware preferences were developed based on two different assumptions, viz., context independent and context dependent assumptions that were appropriate for distinct application scenarios. Eventually, numerous experimentations on real-world dataset revealed that both ways were effective and provided good accuracy against mining personal context-aware preferences for mobile users.

An app recommender approach named Appjoy, dependent on client's app utilization records, to assemble the inclination grid as opposed to utilizing express client judgment, was conceived by Yan and Chen in [16]. AppJoy used cooperative filtering for customizing portable application proposals by considering the client's real application use of designs to take care of the sparsity issue of app utilization.

PROPOSED METHODOLOGY

The existing work uses different approaches to detect fraud apps. This paper presents a methodology for detection of fraud apps using app reviews and ratings, which accurately locates the fraud by mining active periods, also called as leading sessions. Figure 1 shows step by step flow of proposed methodology to deal with fraud detection, which is initiated with reading app reviews and ratings. The .csv extension file of app review dataset is used. This file consists of app reviews, ratings and package name i.e., name of the app. The .csv file provides the number of reviews and ratings of the specific app. Using .csv file the rating is converted to session-based rating by dividing app’s ratings into number of sessions. The session will be as specified by the user. For example, if an app contains 30 reviews and ratings, and session is decided to be of 5, then reviews and ratings distribution will divide in 6 sessions, such as 1-5 reviews and ratings in session 1, 6-10 reviews and ratings in session 2, 11-15 reviews and ratings in session 3, and so on.

Mean rating will calculate using the formula for every session as,

$$\bar{R}_s = \frac{\sum_{i=1}^{M} \frac{\bar{R}_o}{N}}{\sum_{i=1}^{M} \frac{\bar{R}_o}{N}}$$

such that every session will have their mean rating.
Furthermore, two techniques of NLP will be used to detect action words from the reviews i.e., POS tagging and chunking. However, POS tagging will help to mark word in a sentence based on its relationship with related words and chunking will help to get more general or more specific information. After applying NLP, fuzzy logic will help to classify the reviews in more detailed way. Fuzzy logic will categorize reviews into five classes namely; bad, very bad, neutral, good and very good. Thus, categorization of reviews will be improved. This will further help to convert reviews into rating using review analysis.

Fig 1: Workflow of proposed methodology
Next, an approach is proposed to apply session-based analysis on both reviews and ratings, lastly pattern analysis will help to analyze the results and make decisions. This will result in improving differentiation between the reviews and provide better classification accuracy.

CONCLUSION

Many fraud detection techniques have been proposed in recent years. A concise survey on the fraud mobile app detection methods is explained in section II that may benefit researchers to explore new ideas and provide new solutions. This paper gives an overview of the fraud mobile app detection. This paper proposes a new approach for detection of fraud mobile app. It is expected that the approach proposed for fraud mobile app detection should provide better accurate results to detect the fraud apps. It should be less complex but at the same time it should not compromise with the detection accuracy.

REFERENCES

[18] The dataset mentioned in the paper is freely available to reader at the address [online]. Available: https://larc.smu.edu.sg/buzzcity-mobile-advertisement-dataset