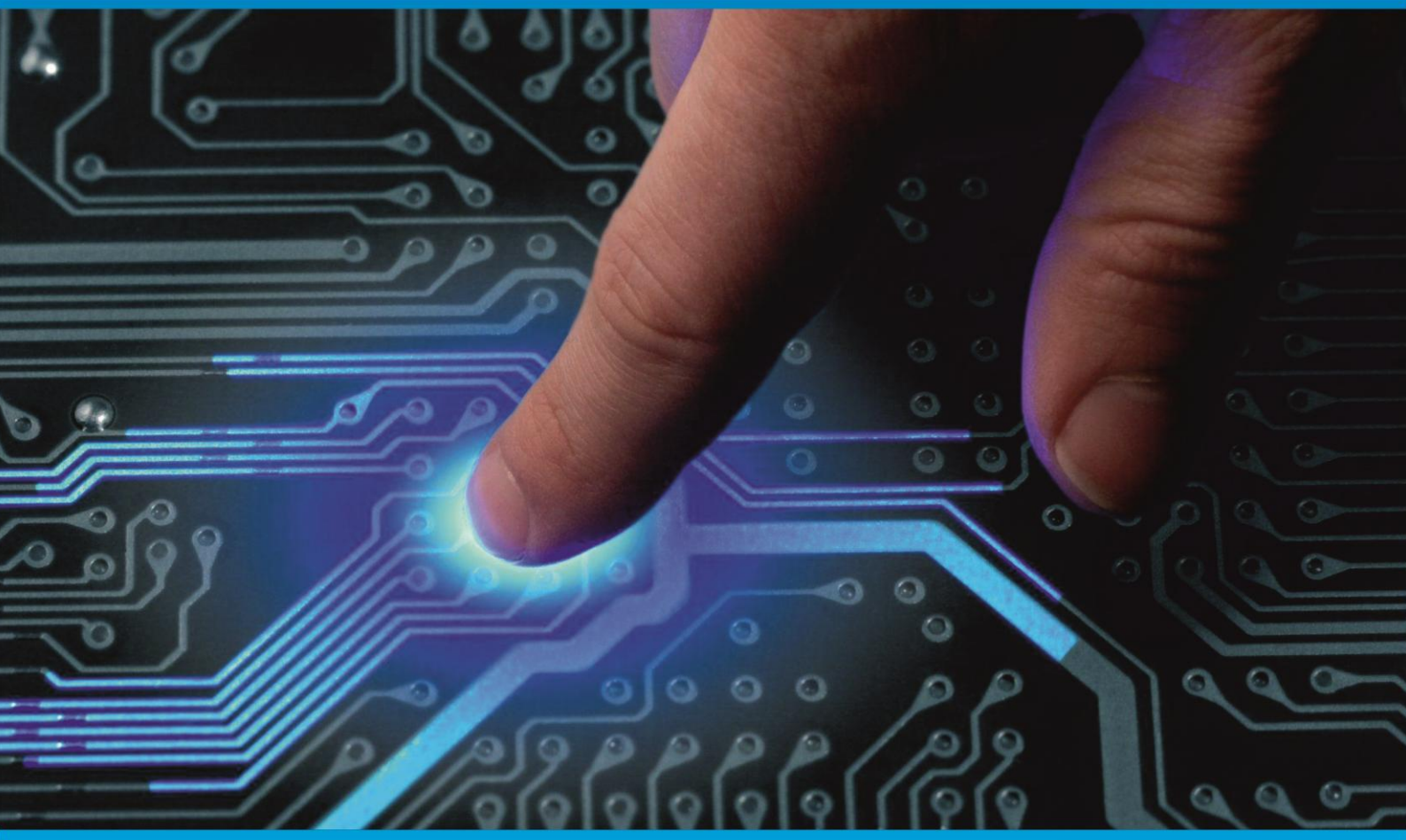
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

Discovery of Ranking Fraud for Mobile Apps

Khan Mohammed Junaid Rafique, Ghembad Pratik Balasaheb, More Aditya Avinash, Shaikh Mohamed Yaseen Faiyyaz, Prof.N.Sharma

BE Students, Department of Computer Engineering, Sinhgad Academy of Engineering, Kondhwa, Pune, Maharashtra, India

Department of Computer Engineering, Sinhgad Academy of Engineering, Kondhwa, Pune, Maharashtra, India

ABSTRACT: 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.

KEYWORDS*:* Sentiment Analysis, Reviews Records, Mobile Fraud detection, Machine learning, Supervised Learning

I. INTRODUCTION

In the previous not many years the quantity of versatile applications develops in profoundly gradual way. Apple's application store and Google play store contains numerous applications. Leaderboard is utilized to show the diagram rankings of most well known applications. For the advancement of versatile applications App Leaderboard is utilized. The application having higher position in leaderboard prompts enormous measure of downloading. In this way, the application engineer procures the different approaches to positioned high their applications in the leaderboard for example advertising.Some application designers utilized the deceitful method to promotetheir applications. They can control the diagram rankings in the application store. Unexpected increment the application downloads, appraisals and surveys are actualized by using"bot ranches" and "human water armed forces". As indicated by article from VentureBeat [1], when an application is advanced by utilizing fake way the positioning is expanded from 1800 to top 25 and in excess of 50,000 100,000 new clients could be gained in two or three days.This positioning extortion sway on portable application industry in extremely huge concern.

The writing work worry about versatile application recommendation [6],[9], online survey spam detection [2] and web positioning spam detection [8],[10]. To conquer the issue of positioning misrepresentation proposed positioning extortion recognition framework. There are a few difficulties to accomplish this. First is that positioning misrepresentation isn't generally occur so we need to distinguish precise timing. Second challenge is that the quantity of portable applications is colossal so physically positioning every single application is troublesome. The applications are positioned high just in their driving meeting which is an assortment of driving occasions. So to recognize the main meetings of each application dependent on its authentic positioning records proposed calculation. At the point when the versatile applications advanced by using fraudulentway, specific design is watched. A few confirmations are separated by contrasting this example and ordinary applications.

To separate just positioning based confirmations isn't adequate so we can extricate rating and survey proves moreover. We utilized proof accumulation technique for the assortment of these three sorts of confirmations. All the confirmations are heterogeneous. So the conglomeration of this confirmations is very testing. All the proof removed are heterogeneous so the displaying of the confirmations are significant. For the displaying of this confirmations measurable speculation test are utilized. For the identification of audit confirmations KMP calculation is proposed[11]. The proposed system is versatile and can be reached out with other space created confirmations for positioning misrepresentation location.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

II. MOTIVATION

- To rank extortion for portable application.
- To improve the extortion identification effectiveness.
- We should initially examine the essential attributes of driving occasions for removing extortion confirmations.
- The suspicious driving occasions may contain exceptionally short rising and downturn stages.
- We ought to break down web positioning spam recognition. In particular, the web positioning spam alludes to any think activities which bring to chosen site pages an unmerited good significance or significance.
- We concentrated on distinguishing on the web survey spam.

III. REVIEW OF LITERATURE

1. In this paper, the author aim to give overview of existing detection approaches in a systematic way, define key research issues, and articulate future research challenges and opportunities for review spam detection. Opinion spam (or fake review) detection has attracted significant research attention in recent years; the problem is far from solved. In this survey, the author presents various methods of opinion spam detection. Further work needs to be conducted to establish how many features are required and what types of features are the most beneficial.
2. This survey has explored almost all published fraud detection studies. It defines the adversary, the types and subtypes of fraud, the technical nature of data, performance metrics, and the methods and techniques. After identifying the limitations in methods and techniques of fraud detection, this paper shows that this field can benefit from other related fields. Within the business context of mining the data to achieve higher cost savings, this research presents methods and techniques together with their problems. Compared to all related reviews on fraud detection, this survey covers much more technical articles and is the only one, to the best of our knowledge, which proposes alternative data and solutions from related domains. Future work will be in the form of credit application fraud detection.
3. Click fraud represents a serious drain on advertising budgets and can seriously harm the viability of the internet advertising market. This paper proposes a novel framework for prediction of click fraud in mobile advertising which consists of feature selection using Recursive Feature Elimination and classification through Hellinger Distance Decision Tree. This paper has developed a novel framework to detect fraudulent partners based on click data associated with mobile phone internet surfing .New features based on the attributes were generated, and these features were used to model the behavior of each partner's click. Further the performance will be improved by using various techniques.
4. The need for paying with mobile devices has urged the development of payment systems for mobile electronic commerce. Most of the available fraud and intrusion detection systems for e-payments are specific to the systems where they have been incorporated. This paper proposes a generic model called as Activity Event-Symptoms model for detecting fraud and intrusion attacks which appears during payment process in the mobile commerce environment. The proposed scheme identifies the intrusions/frauds happening in customer accounts and vendor accounts by identifying the various suspicious symptoms in business transactions. The system emphasizes on-line analysis of transactions instead of offline analysis.
5. In the paper, web spam has been considered as a crucial challenge in the world of searching. We explained various methods of web spamming and algorithms to combat with web spam. Up to now, many methods have been created to combat with web spam. However, due to its economical profit and attractiveness, on one side, researchers have presented new methods to combat with it, and in another side, spammers present some methods to overcome these limitations. We hope that we can observe spam pages reduction by presenting character algorithms to detect web spams.
6. This paper research focuses on systematically analyzing and categorizing models that detect review spam. Try to present on a organized review of web spam detection techniques with the emphasis on algorithms and underlying principles. Categorize all existing algorithms into three categories based on the type of information they use i.e. is content based methods, link-based methods, and additional methods based on non-traditional data based on the user behavior with the different sessions. In addition, machine learning can be used to contribute in discovering web spam pages.
7. In this paper, we develop ranking fraud detection system for mobile apps. It reviews various existing strategies used for internet or web spam detection, which is associated with the rating fraud for mobile Apps. Also,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

we've seen references for online review unsolicited mail detection and mobile App advice. By using mining the main sessions of mobile Apps, we aim to locate the ranking fraud. The leading classes works for detecting the nearby anomaly of App ratings. The machine targets to locate the ranking frauds based on three styles of evidences, including rating based evidences, ranking based evidences and comment based evidences. In addition, an optimization based totally aggregation method combines all of the three evidences to hit upon the fraud.

8. Credit card fraud is escalating significantly with the advancement of modernized technology and became an easy target for frauds. Credit card fraud has highly imbalanced publicly available datasets. In this paper, the authors apply many supervised machine learning algorithms to detect credit card fraudulent transactions using a real-world dataset. This system identifies the most important variables that may lead to higher accuracy in credit card fraudulent transaction detection. Additionally, they compare and discuss the performance of various supervised machine learning algorithms that exist in literature against the super classifier that we implemented in this paper. Furthermore, we employ these algorithms to implement a super classifier using ensemble learning methods.
9. The insurance industries consist of more than thousand companies in worldwide. And collect more than one trillions of dollars premiums in each year. When a person or entity make false insurance claims in order to obtain compensation or benefits to which they are not entitled is known as an insurance fraud. The traditional approach for fraud detection is based on developing heuristics around fraud indicator. The auto\vehicle insurance fraud is the most prominent type of insurance fraud, which can be done by fake accident claim. In this paper, focusing on detecting the auto\vehicle fraud by using, machine learning technique. In this paper, just bring out the feature of machine learning algorithms. In future can be work with more algorithms and calculate which provide more accuracy, precision, and recall.
10. This paper discusses the commonly used supervised algorithms. The primary goal was to prepare a comprehensive review of the key ideas and present different techniques for every supervised learning method. The paper makes it clear that every algorithm differs according to area of application and no algorithm is more powerful than the other in different scenarios. The choice of an algorithm should be made depending on the type of problem given to us and the data available. The accuracy can be increased by using two or more algorithm together in suitable conditions

IV. SYSTEM OVERVIEW

A novel proposed framework is to representative a given review dataset as a Heterogeneous Information Network (HIN) and to solve the issue of spam detection into a HIN classification issue. In particular, to show the review dataset as a HIN in which reviews are connected through different node types (such as features and users). A weighting algorithm is then employed to calculate each feature's importance (or weight). These weights are applied to calculate the final labels for reviews using both unsupervised and supervised methods. Based on our observations, defining two views for features (review-user and behavioral-linguistic), the classified features as review behavioral have more weights and yield better performance on spotting spam reviews in both semi-supervised and unsupervised approaches. The feature weights can be added or removed for labeling and hence time complexity can be scaled for a specific level of accuracy. Categorizing features in four major categories (review-behavioral, user-behavioral, review-linguistic, user-linguistic), helps us to understand how much each category of features is contributed to spam detection.

1. NetSpam framework that is a novel network based approach which models review networks as heterogeneous information networks.
2. A new weighting method for spam features is proposed to determine the relative importance of each feature and shows how effective each of features are in identifying spams from normal reviews.
3. NetSpam framework increases the accuracy as opposed to the state-of-the art in terms of time complexity, which distinctlyrelies upon to the variety of capabilities used to perceive an unsolicited spam evaluation.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

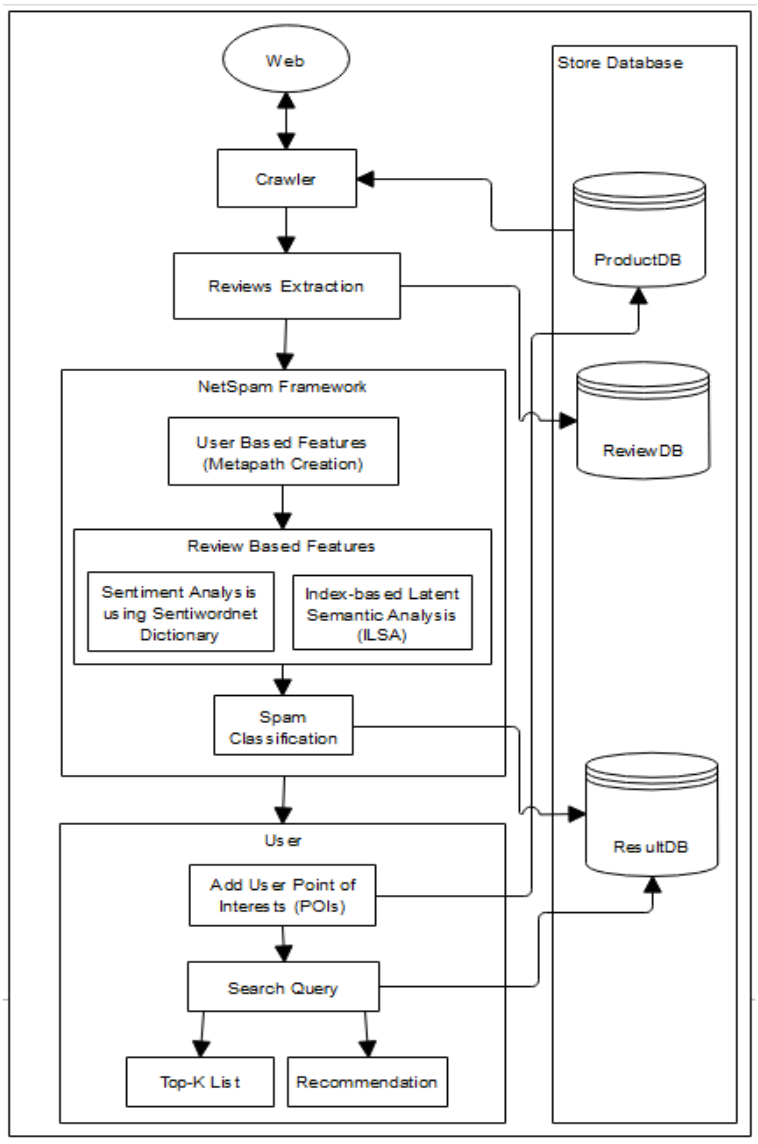


Fig.1 Proposed System Architecture

The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network and to map the problem of spam detection into a HIN classification problem. In particular, model review dataset as in which reviews are connected through different node types. The fig. 2 shows the flowchart of NetSpam framework.

A weighting algorithm is then employed to calculate each feature's importance. These weights are applied to calculate the final labels for reviews using both unsupervised and supervised techniques. Based on the observations defining two views for features.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

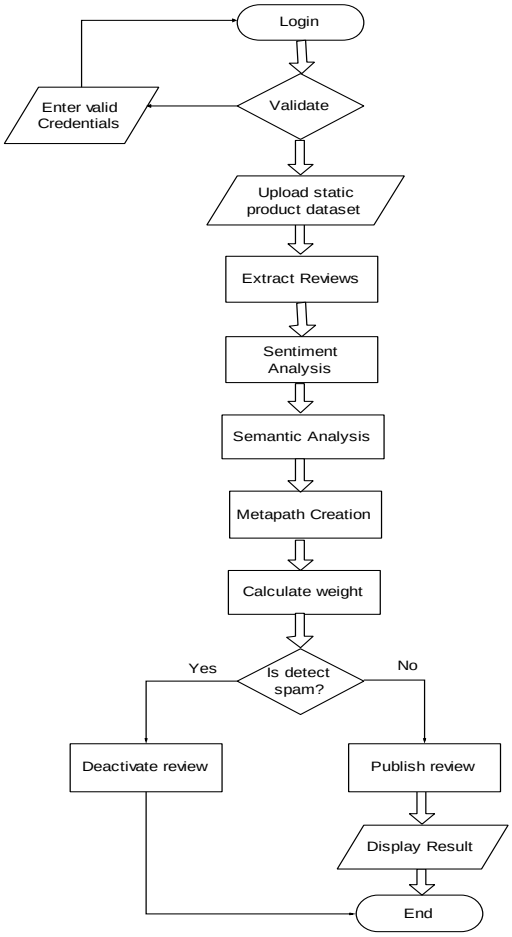


Fig. 2 Flowchart of NetSpam Framework

Advantages:

1. To identify spam and spammers as well as different type of analysis on this topic.
2. Written reviews also help service providers to enhance the quality of their products and services.
3. To identify the spam user using positive and negative reviews in online social media.
4. To display only trusted reviews to the users.

V. FEATURES

This paper use metapath concept to establish link between reviews as follows. A metapath is defined as a path between two reviews, which indicates the connection of two reviews through their shared features. When talk about metadata, refer to its general definition, which is data about data. In our case, the data is the written review, and by metadata mean data about the reviews, including user who wrote the review, the business that the review is written for, rating value of the review, date of written review and finally its label as spam or genuine review.

Metapath is created using following features:-

A. User Behavioral

These features are related to each individual user and they are calculated per user. Thus use these features to generalize all the reviews written by that specific user. This category has two main features- Burstiness of reviews written by a single user and the average of a users' negative ratio given to different businesses.

Burstiness- Spammers, generally write their spam reviews in short time as they want to impact users, and since they are temporal users.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

Negative ratio- Spammers usually write reviews which defame businesses which are competitors to the ones they have contract with. This is done with destructive reviews, or by rating such businesses with low scores. Hence, ratio of their result tends to be low. Users with average score equal to 2 or 1 take 1 and others take 0.

B. User Linguistic

These features are extracted from the users' language and show how users are describing their feelings or opinions about what their experiences were being a customer of a particular business. There are two features intended for our framework in this category; Average Content Similarity (ACS) and Maximum Content Similarity (MCS).

Average Content Similarity and Maximum Content Similarity- Spammers, generally write their reviews with same predefined template and they generally prefer not to waste their time to write an original review. As a result, they have same reviews. Users have close calculated values take same values (in [0, 1]). This feature requires semantic analysis to be performed to detect copy paste mechanisms used by spammers. Then detect copy paste employed by spammers by calculating time between the start of writing their fake reviews and submitting their reviews. Copy paste employed requires less time to post a review of many words than the time required to actually write the same review manually.

C. Review Behavioral

This feature type is based on metadata of review and not on the review text itself. The RB category consists of two features; Early time frame and Threshold rating deviation of review.

Early Time Frame - Spammers try to write their reviews as soon as possible, so as to keep their reviews in the top reviews.

Rate Deviation-: Spammers, try to enhance businesses they have settlement with, so they rate these businesses with very high scores. In outcome, there is high diversity in their given scores to various types of businesses which is the reason they have high variance and deviation. Average of the review ratings change drastically over a week's time which can be detected using rate deviation and calculation of entropy.

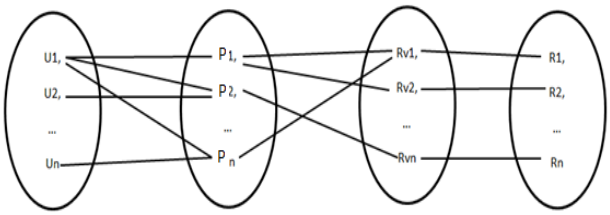
D. Review Linguistic

This feature is based on the review itself and extracted directly from text of the written review. In this work to utilize two main features of RL category; the Ratio of 1st Personal Pronouns (PP1) and the Ratio of exclamation sentences containing '!'.Study shows that spammers use second person pronouns often and use more of exclamation marks to create an impression on readers.

Reviews are similar to each other based on their calculated value, take same values (in [0, 1)).

V. MATHEMATICAL MODULE

A. Mapping Diagram



Where,

U1; _ _ _;Un = No. of Users

P1; _ _ _;Pn = List of Products

Rv1; _ _ _;Rvn = List of reviews per product

R1; _ _ _;Rn = Spam detection result

B. Set Theory

Let us consider S as a set of NetSpam framework

S= { }

INPUT:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905253 |

- Identify the inputs as low resolution images
  F= {f_1, f_2, f_3 .....,f_n| 'F' as set of functions to execute to spam detection framework}
  I= {i_1, i_2, i_3…|'I' sets of inputs to the function set}
  O= {o_1, o_2, o_3….|'O' Set of outcomes from the function sets}
  S= {I, F, O}
  I = {user metadata and his/her feedback/review}
  O = {Classification of spam detection result}
  F = {metapath creation, sentiment analysis, semantic analysis and calculate weight}

Space Complexity:
The space complexity depends on project presentation and visualization of discovered patterns. More the storage of data reviews more is the space complexity.
Time Complexity:
Check No. of patterns available in the datasets= n
If (n>1) then retrieving of information may be time consuming. So the time complexity of this algorithm is O(n^n).
Φ = Failures and Success conditions.
Failures:
1. Huge review database can lead to more time consumption to get the information.
2. Hardware failure.
3. Software failure.
Success:
1. Search the required information from available in Database.
2. User gets result very fast according to their needs and point of interests.

VII. CONCLUSION

Work can built a ranking fraud detection system for mobile Apps. To built such a system , first showed that ranking fraud detected in leading sessions and provided a algorithm for mining leading sessions for each App from its historical ranking records. Then, extract ranking based evidences, rating based evidences and review based evidences from the historical records to detect ranking fraud. To overcome this fraud proposed an optimization based aggregation method to integrate all the evidences for estimating the credibility of leading sessions from mobile Apps. A different perspective of this approach is that all the evidences can be modeled by statistical hypothesis tests, thus it is easy to be enlarged with other evidences from domain knowledge to detect ranking fraud. Finally, validate the proposed system with extensive experiments on real-world App data collected from the Google Play store.

REFERENCES

[1] Hengshu Zhu, HuiXiong, Yong Ge, Enhong Chen, "Discovery of ranking frauds for mobile apps,", IEEE transactions on knowledge and data engineering, vol. 27, no. 1, january 2015.
[2] Ee-Peng Lim, Viet-An Nguyen,Nitin Jindal, Bing Liu, Hady W. Lauw, "Detecting Product Review Spammers using Rating Behaviors", CIKM10, October 2630, 2010.
[3] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, "taxi driving fraud detection system", in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 181190.
[4] N. Jindal and B. Liu, "Opinion spam and analysis,"in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219230.
[5] A. Klementiev, D. Roth, and K. Small, "An unsupervised learning algorithm for rank aggregation,",in Proc. 18th Eur. Conf. Mach. Learn., 2007, pp. 616623.
[6] H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, and J. Tian, "Mining personal context-aware preferences for mobile users,", in Proc. IEEE 12th Int. Conf. Data Mining, 2012, pp. 12121217.
[7] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery,", in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 823831.
[8] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis,", in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 8392.
[9] K. Shi and K. Ali, "Getjar mobile application recommendations with very sparse datasets,", , in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 204212.
[10] N. Spirin and J. Han, "Survey on web spam detection: Principles and algorithms,", SIGKDD Explor. Newslett., vol. 13, no. 2, pp. 50 64, May 2012.

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details