iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41511

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Fake Rating Analysis from the Movie Dataset: Using Supervised Techniques

Ms. Nirmal Kaur
Research Scholar1, Department of Computer Science and Application, Sant Baba Bhag Singh University

Abstract: ***A social networking site provides a platform for different users to interact with. This forum attracts not only people but also business owners by selling products. The most commonly used social media platform is to promote online films. For this purpose, fake profiles are created by developers. Finding such a false profile to reveal actual movie ratings is the main objective of this study. For fake profile categories such as pre-processing, classification and classification are followed. In pre-processing, sound or other impurities are removed from the movie database if it contains. The next step is to classify the features and select the functional features based on the analytical analysis. Separation uses a Raphson-based repetitive model in testing a fake profile. The findings of the proposed method are presented in terms of F-Score, category accuracy, sensitivity and clarity. The overall result is an improvement in important genes.***

Keywords: Fake Rating, Supervised Techniques, F-score, Classification Accuracy

I. INTRODUCTION

Nowadays, fake profiles are everywhere online. These profiles may be found on social media accounts. [1] To improve ecommerce sites, it is mandatory to obtain a fake profile that helps us find useful products on ecommerce sites. Proper marketing of products will lead to improved product sales that automatically increase company profits. To increase company sales, they use any means whether they are right or wrong. [2] The proposed system assists in detecting attacks and the fake profile provided for the development of movies. [3]The promoters promote the movies and enhance their sales of movies ticket by adding fake reviews of the movies. The methodology tht is used to detect the fake profiles contains the following phases. The phase or step is pre-processing which is done to remove noise or impurities from the dataset. [4]The dataset used for the proposed work is taken from movie lense. [5][6]Like other datasets, the movie lense dataset also contains the impurities and noise like missing values. Moreover, it also contains some string literals. The classification accuracy is enhanced by converting these literals into nominal form. [7][8] This can be done with the help of nominal conversion and correlation that helps in performing overall operation. Another phase is feature extraction which aids in extracting the features which have highest correlation among other dataset. [9], [10][11] [12], [13] After that Feature selection techniques are employed in order to select the required features. For bounding the profiles into the clusters and determining the fake profiles, one of the popular method known as Raphson method is used which finds the false or fake profiles with the help of IP address of the fake reviewer. For classifying the clusters KNN based Raphson technique is used which helps in producing the best result in the forms of the classification accuracy. The second section of the paper gives gives the information related to the contribution of the work done by other authors which is known as literature review or survey. Similarly, section 3 provides the information about the proposed methodology which is adopted to achieve the objectives, the section 4 provides the information regarding the performance analysis of the system along with the results obtained from the thesis , whereas section 5 provides the conclusion and future work of the proposed system, the end section of this paper gives the references detail.

II. LITERATURE SURVEY

The literature survey section highlights the some of the work done by the other authors in the field of the proposed system. They have done their work on the social datasets in order to find the false profiles. One of the best approaches which is used to detect the false profiles is Data mining. This section contains the review of the papers that are published in the Springer, IEEE and other reputed websites. [14]demonstrate an identity detection mechanism which is multiple in natures from the various social networking sites dataset. For detecting, they employed behavior detection process. They examined physical characteristics at initial step. They lag behind the classification accuracy. However, their detection of false profile was accurate. [15]illustrated that social media dataset uses clustering procedures to identify fake profile. To achieve their task, they used social media dataset. The main demerit of their system was the inability to handle the missing values which lead to low classification accuracy. [16]used a twitter dataset for detecting false profile. To identify abnormal patterns, they used pattern recognition techniques. As they have not done the preprocessing on their dataset, this proves a little loop hole for their research and the accuracy of their system has to suffer due to

this pit hole. [17]implemented graph-based approach to detect the false in the social networking dataset. To detect the fake profiles they uses the concept of cycles which is found in the Graph based approach. The next node is discovered with the help of path traversing techniques present in the graphs. When the same path was followed again and again, this will help in finding the false profile. They represented their result by finding the classification accuracy of the proposed system. [18]created a system which detects the fake profiles on the online social networking. The abnormal patterns were discovered with the help of dataset. The dataset is taken from the one of the popular website called Kaggle. They also represented their result in the terms of classification accuracy. [19]demonstrated the concept of clone attack using social media dataset. The term clone attack means that the same sender sends the infected data to the destination again and again which threats the data. For solving this problem, a detection technique was discovered which was based on mining. Like other system, they also used classification accuracy term to check the efficiency of the system.[20]demonstrated cross domain attribution based on authorship that test fake profile based on gender and bot. They firstly use prepossessing mechanism to remove noise from the dataset. After that for classification and segmentation was done by the algorithm based on Machine learning. They did not evaluate the performance of the system. [21]used the bot method to detect false profiles in the dataset of the online social media. The impurities and extra fields were removed by correlation-based technique from the dataset. They used mining-based techniques to detect the false profiles. The better results were given by the positive correlation technique and efficiency of the system was given by the classification accuracy. [22]used another technique called artificial intelligence that detects the bots in dataset that was taken from social media. They used learning mechanism to detect fake profiles which was supervised in nature. The efficiency of the system is expressed in terms of sensitivity, classification accuracy and specificity. [23]demonstrated procedure that detects false profiles on the social media. They detected the false profiles by mining-based approaches which are based on emotions. Like other system they used pre-processing techniques for tackling the problem of noise. The results showed that the accuracy of the proposed system is very high.

The following figure 1 showed that majority of the papers studied are from IEEE access and the least contribution is shared by ACM.

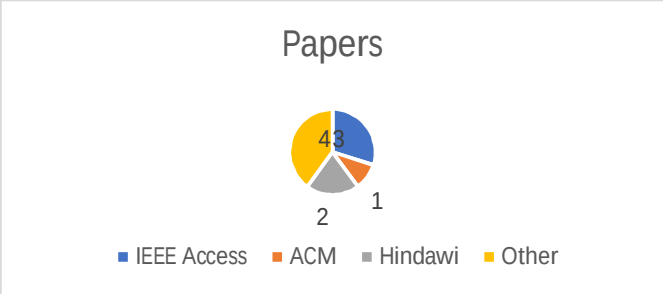


Figure 1: Papers categorization

III. RESEARCH METHODOLOGY

The proposed methodology is classified into sections or phases. The step is of preprocessing in which noise is removed. The pre-processing mechanism uses correlation-based attribute elimination. In addition, noisy values are eliminated using mod based approach. In the second phase, segmentation is applied. Segmentation partitioned the entire dataset into critical and non-critical parts. Segmentation is implemented through similarity based KNN approach. Feature extraction with Iterative Raphson based mechanism and The classification is done to finalize the result in terms of identification of attacking nodes. The classification was tested using KNN ,CNN and hybrid approach of KNN with Raphson approach. The detailed methodology is given as under

The pre-processing mechanism used to remove noise from the dataset. This implication of this step is critical since classification accuracy greatly enhanced. The correlation is evaluated with the class field within the dataset with each and every attribute and highest correlation attributes are retained. Correlation is evaluated using the following equation

$$r = \left(\frac{\sum(x_i - mean(x))(y_i - mean(y))}{\sqrt{\sum(x_i - mean(x))^2 \sum(y_i - mean(y))^2}}\right.$$

Equation 1: Correlation evaluation

Here ‘x’ will be a part of ratings and ‘y’ will be ip_address. Since they present psotive correlation. The range of values taken from field ratings will be added within x variable and range of values from ip_address will be added within y variable.

The string dataset attributes like movie_name ,date and time ,actors, director and lyricist produced negative correlation. These attributes are first of all converted to nominal form. In this case string values are converted to 0. The conversion represents string attributes using set of The correlation values lie between -1 to 1. The original dataset is reduced on the basis of correlation. The negative correlation attributes are eliminated from the original dataset. The noise from the dataset is eliminated using mode based approach. The mode-based approach calculates the highest repetitive values from the attribute and then eliminates the missing values with highest frequency values. This help in increasing classification accuracy.

Segmentation through the Similarity based mechanism The proposed mechanism evaluates the similarity between the user data based on identification number. Each user is also assigned with the unique ipaddress. Using Euclidean based mechanism, clustering is done, after the clustering similar identification-based nodes are labelled as malicious. Rest of the nodes are labelled as fair nodes and does not play a part in identification process. The segmentation process is described in the following figure.

Figure 2 describes the similarity-based clustering is done within segmentation phase. The mechanism is described through the following diagram. Figure 3 describe the cluster-based approach for the proposed approach. Cluster 1 and cluster 2 contains ids that are similar based on attribute ip address. Cluster 3 represents unique ids thus no fake profile exists in this case. Cluster identification numbers are obtained using Raphson method discussed in next section. INTERATIVE RAPHSON METHODfinds the similarity between same profiles-based users and calculates that similarity based on historical rating of similar items-based neighboring given by all users. INTERATIVE RAPHSON METHOD classified user's categories into fake and genuine.

'X' values will correspond to individual tuple or row from pre-processed dataset. The entire information is represented as

1	Wild Love	Drama	02/09/2020	5	30.255.237.145

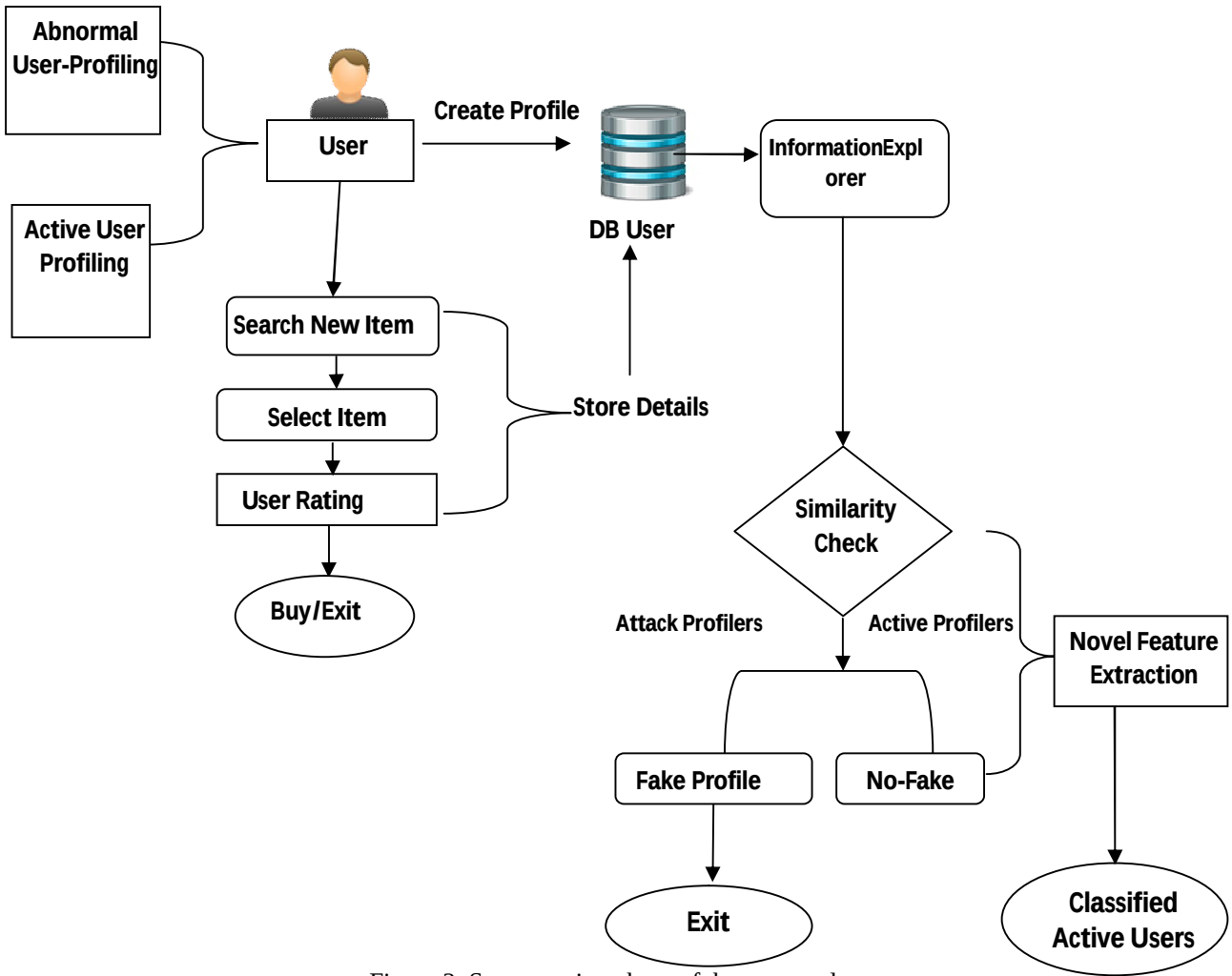


Figure 2: Segmentation phase of the proposed system

Out of this first column is userid , second column is movie name and rest of the attributes are movie type, date of release etc. out of these columns string columns present negative correlation and hence are eliminated. User id, rating and id addressing is retained. These three values will serve as X=[userid, rating, round(floor(ip address))]. These values will be used as roots within the Raphson equation. Value of y will serve as cluster id.

This mechanism is iterative and performs operation by categorizing the dataset into set of independent and dependent variables. The 'x' represents the independent variable whereas 'y' denotes the dependent variable. 'x' values represent user id, ratings , ip address by performing decimal point elimination process. This means from 123.45.67.120 only 123 will be used.

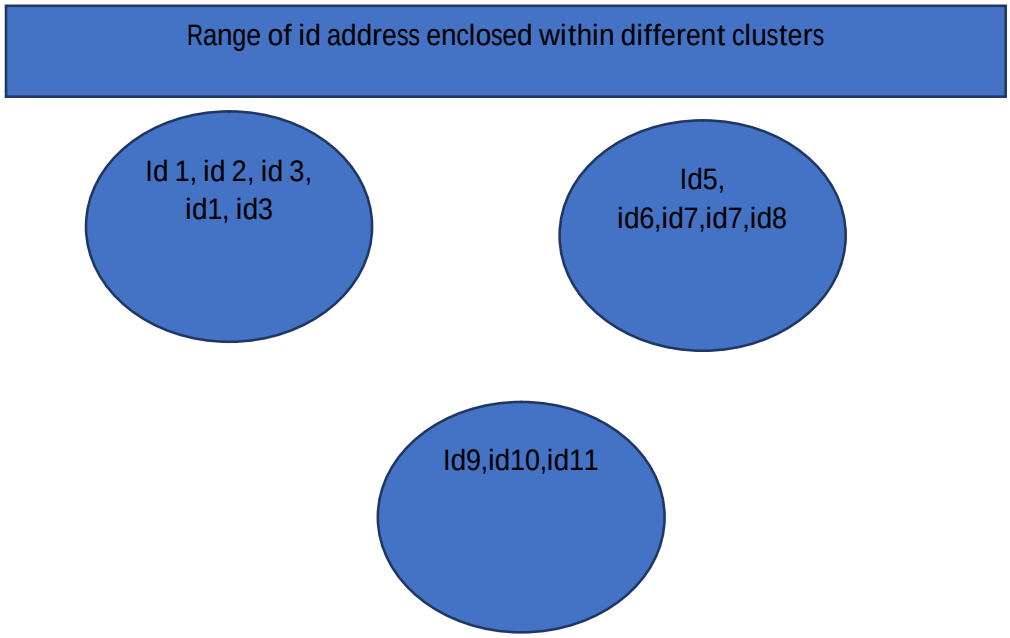


Figure 3: Clustering with the proposed system

Here f(x) represents optimization function represented as

$$f(x) = x^2 - length(dataset)$$

Equation 2: Optimization equation

'x' is the variable extracring the values from highest correlated attribute. The clusters are labelled and in order to identify the labelling for the data item following equation will be used.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Equation 3: Cluster label identification equation

'$X_{n+}1$' is the next root for the equation. The objective function for the same is utilized. This objective function is labelled in the equation 3

$$f(x) = x^2 - n$$

Equation 3: Objective or fitness function

'x' is the variable getting the value from the dataset and 'n' is the total number of attributes existing within dataset. We try to optimize this function to determine label which act as rating given by user. 'n' in our case number of clusters is assumed to be '5'. Hence equation 3.2 will become.

$$f(x) = x^2 - 5$$

Equation 4: Example of objective function

Derivative of above equation is given as

$$f'(x) = 2x$$

Equation 5: Derivative of equation 4

The initial approximation for x1=2.6 is taken and equation 3.1 is used for next approximation.

$$x_1 = 2.6 - \frac{1.76}{5.2}$$

$$x_1 = 2.261$$

Equation 6 : Root from the equation 5

The next iteration is performed again and initial approximation is used as 2.261.

$$x_1 = 2.261 - \frac{0.1121}{4.522}$$

$$x_1 = 2.23$$

Equation 7: root for next iteration from equation 5

Now there is a little difference between the previous label and new label. Thus rating corrected to one decimal place is obtained. This 2.23 value will serve as cluster identification.

The identification process leads to fake user detection in case multiple ratings are generated through the same identification number. The classification phase presents the result of the applied segmentation and feature extraction mechanism.

The classification for the proposed approach is tested using KNN and CNN .The classification results are presented in the form of classification accuracy, sensitivity, specificity and F-Score. The result achieved with the proposed approach is suddenly better as compared existing approach.

IV. PERFORMANCE ANALYSIS AND RESULTS

Performance analysis and result first of all describe the metrics used within the proposed system. The parameters corresponding to proposed approach is given as under

A. Classification Accuracy

The proposed system's accuracy is the highest accuracy when it is compared with other systems. it does not work with vector machines that support multi values. The correct predictions that the system makes from the total prediction is given by classification accuracy. To exemplify, if the system has total 100 values and 89 will be predicted right, then its accuracy is 89%. The equation 8 gives the formula to calculate classification accuracy.

$$CA = \frac{Total_{Correct_{Predictions}}}{Total_{Predictions}}$$

Equation 8: Classification Accuracy

B. Sensitivity

Sensitivity deals with the number of the results which are positive out of the total samples. To cite an example, the soil of the plants in which they grow is measured in sensitivity. This means detecting soil for the plant growth is measured in terms of sensitivity. The formula to calculate sensitivity is given by equation 9.

$$Sensitivity = \frac{Positive_{results}}{Total_{Preidctions}}$$

Equation 9: Sensitivity

C. Specificity

Specificity is the negation of sensitivity. It is the negative prediction that does not provide the results. In other words, it provides negative results to the users. The formula t calculate specificity is given by equation 10.

$$Specificity = \frac{Negitive_{result}}{Total\_Predictions}$$

Equation 10: Specificity

D. F-Score

It is the metric which uses the concept of classification accuracy to check the efficiency of the system. It is derived from negative, true negative, false positive and positive values. The formula to calculate this is given by equation 11.

$$F - Score = \frac{TP}{TP + \frac{1}{2(FP+FN)}}$$

Equation 11: F-score

The recommendation of an item to an active user (UA) is determined by the prediction distance computed. To recommend a set of N items to the active user (UA), we firstly form a group of similar profiles to the active user based on the similarity of using approach. Check the relation between item and user compared with based on user historical search. The classification accuracy from the proposed approach is given as under

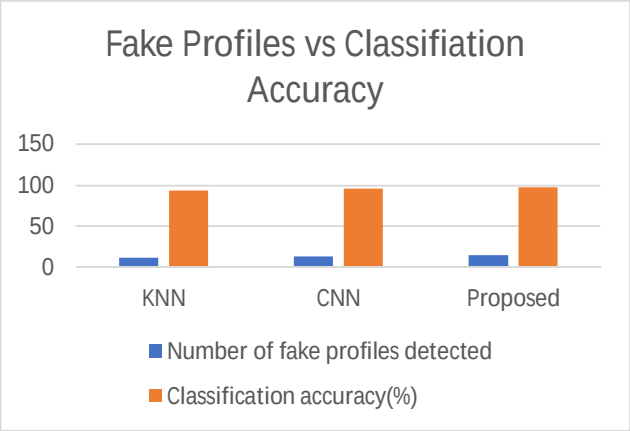


Figure 4: Classification Accuracy

The proposed approach classification accuracy in the range of 98% which is significantly higher as compared to KNN and CNN.

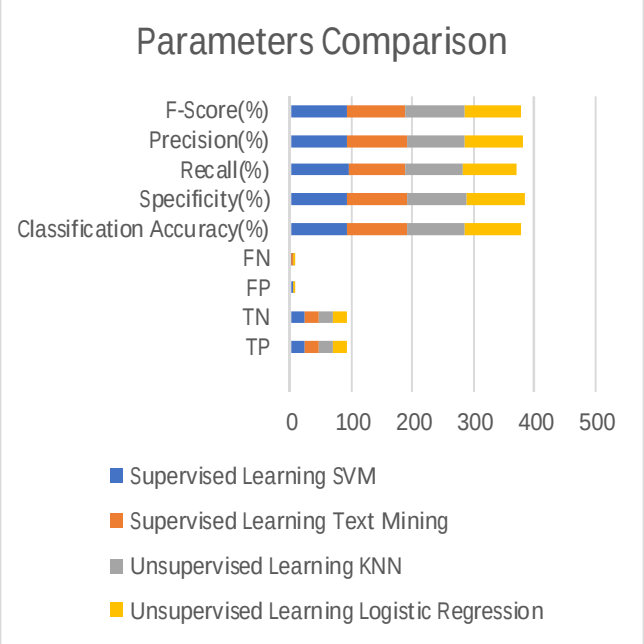


Figure 4: Comparison of supervised and unsupervised learning

The comparison of supervised and unsupervised learning is made to determine the best possible approach to be used for fake profile detection.

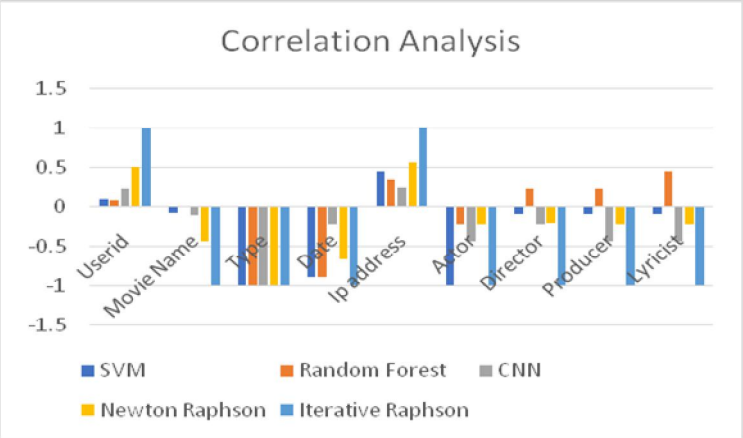
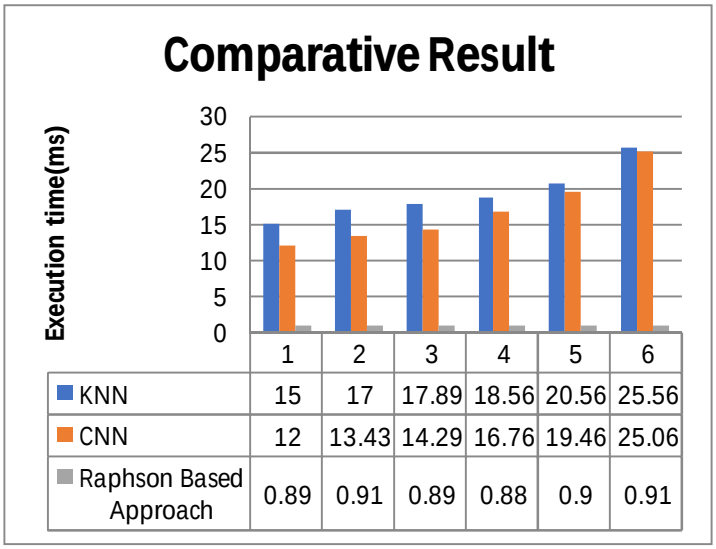


Figure 5: Correlation analysis

This comparison is made to determine the approach that yield absolute values for useful features from the dataset. Iterative Raphson approach produces better result as compared to other approaches.



	1	2	3	4	5	6
KNN	15	17	17.89	18.56	20.56	25.56
CNN	12	13.43	14.29	16.76	19.46	25.06
Raphson Based Approach	0.89	0.91	0.89	0.88	0.9	0.91

Figure 6: Execution time

The proposed approach proves to provide quick result. The execution time consumed for the proposed work is much lesser as compared to existing approach.

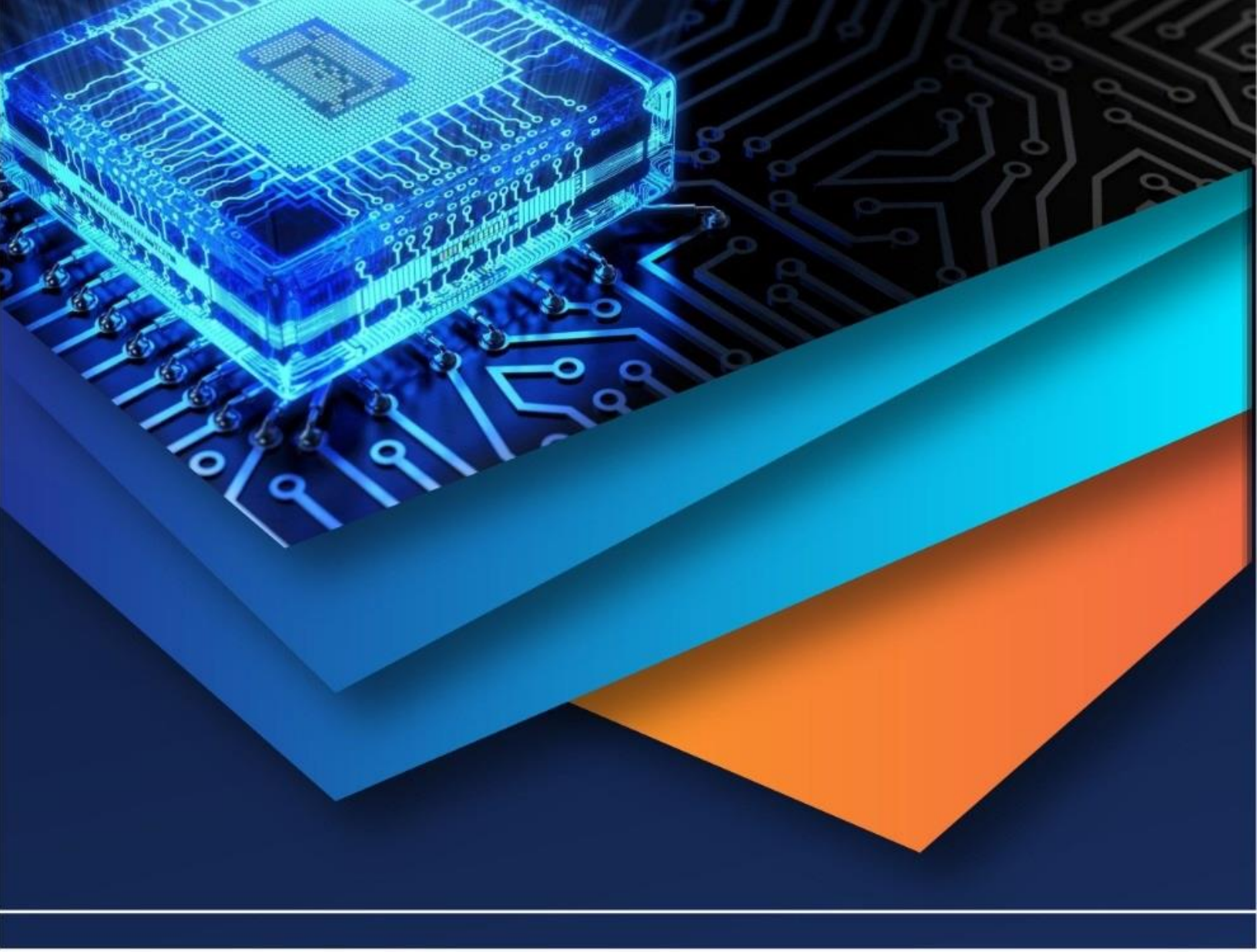
This algorithm searches for innovators similar to the target, it is very probable that the recommended item would be similar the target's profile, which is a disadvantage. Whereas the advantage of this technique is that is takes into consideration the release time of an item and the consumption time too. It recommends unpopular items, since innovator probability gives a higher score to items that have been recently released and are not yet popular.

V. CONCLUSION AND FUTURE SCOPE

This operation uses the steps to achieve the detection of a fake profile early. The hacked system is injected with invasive nodes. The pre-processing phase is used to remove audio when present within the database. Audio is removed using word modification and mode-based mode. The purpose of this method is to increase the accuracy of the sections. The next step is the classification section used to extract logical attributes from the database. In the last stage a split is made. Comparisons of different approaches were made to prove the importance of the study. Using the benchmark database found in the film lens. The process of identifying a fake profile is developed using the proposed method. The result shows an improvement in important genes. The future scope of this study is to include the optimized DDA approach and examined the approach with the real time dataset.

REFERENCES

[1] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in facebook," Jul. 2017, doi: 10.1109/ISEASP.2017.7976996.
[2] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network sybils in the wild," ACM Trans. Knowl. Discov. Data, vol. 8, no. 1, 2014, doi: 10.1145/2556609.
[3] M. Ahmad Wani and S. Jabin, "A sneak into the Devil's Colony-Fake Profiles in Online Social Networks," 2017. Accessed: Feb. 21, 2021. [Online]. Available: https://www.facebook.com/legal/terms.
[4] N. Jindal, B. Liu, and E. P. Lim, "Finding unusual review patterns using unexpected rules," in International Conference on Information and Knowledge Management, Proceedings, 2010, pp. 1549–1552, doi: 10.1145/1871437.1871669.
[5] B. Wu, V. Goel, and B. D. Davison, "Topical TrustRank: Using topicality to combat web spam," in Proceedings of the 15th International Conference on World Wide Web, 2006, pp. 63–72, doi: 10.1145/1135777.1135792.
[6] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in Proceedings of the 15th International Conference on World Wide Web, 2006, pp. 83–92, doi: 10.1145/1135777.1135794.
[7] H. J. Kim, D. K. Chae, S. W. Kim, and J. Lee, "Analyzing crowdsourced promotion effects in online social networks," in Proceedings of the ACM Symposium on Applied Computing, Apr. 2016, vol. 04-08-April-2016, pp. 820–823, doi: 10.1145/2851613.2852005.
[8] H. Jamil, S. Kramer, and R. Wong, "Session details: Volume I: Artificial intelligence and agents, distributed systems, and information systems: Data mining track," Apr. 2016, doi: 10.1145/3252791.
[9] Y. Elyusufi, Z. Elyusufi, and M. A. Kbir, "Social networks fake profiles detection based on account setting and activity," in ACM International Conference Proceeding Series, Oct. 2019, pp. 1–5, doi: 10.1145/3368756.3369015.
[10] J. Kawtar and M. Tomader, "Comparative study about the routing protocols on the vehicular networks and the v2v communications," Oct. 2019, doi: 10.1145/3368756.3369014.
[11] B. Oumayma, "Social media made me buy it: The impact of social media on consumer purchase behavior," Oct. 2019, doi: 10.1145/3368756.3369016.
[12] T. Y. Liu, "Learning to rank for Information Retrieval," Found. Trends Inf. Retr., vol. 3, no. 3, pp. 225–231, 2009, doi: 10.1561/1500000016.
[13] C. Castillo and B. D. Davison, "Adversarial Web search," Found. Trends Inf. Retr., vol. 4, no. 5, pp. 377–486, 2010, doi: 10.1561/1500000021.
[14] M. Tsikerdekis and S. Zeadally, "Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behavior," IEEE Trans. Inf. Forensics Secur., vol. 9, no. 8, pp. 1311–1321, Aug. 2014, doi: 10.1109/TIFS.2014.2332820.
[15] D. M. Freeman and T. Hwa, "Detecting Clusters of Fake Accounts in Online Social Networks Categories and Subject Descriptors," IEEE Access, 2015.
[16] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in ACM International Conference Proceeding Series, Jul. 2015, vol. 2015-July, doi: 10.1145/2789187.2789206.
[17] M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms," Secur. Commun. Networks, vol. 2018, 2018, doi: 10.1155/2018/5923156.
[18] Y. Boshmaf et al., "Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs," Comput. Secur., vol. 61, pp. 142–168, Aug. 2016, doi: 10.1016/j.cose.2016.05.005.
[19] L. Jin, H. Takabi, and J. B. D. Joshi, "Towards active detection of identity clone attacks on online social networks," in CODASPY'11 - Proceedings of the 1st ACM Conference on Data and Application Security and Privacy, 2011, pp. 27–38, doi: 10.1145/1943513.1943520.
[20] W. Daelemans et al., "Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11696 LNCS, pp. 402–416, doi: 10.1007/978-3-030-28577-7_30.
[21] M. Stella, E. Ferrara, and M. De Domenico, "Bots increase exposure to negative and inflammatory content in online social systems," Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 49, pp. 12435–12440, Dec. 2018, doi: 10.1073/pnas.1803470115.
[22] K. C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," Hum. Behav. Emerg. Technol., vol. 1, no. 1, pp. 48–61, Jan. 2019, doi: 10.1002/hbe2.115.
[23] B. Ghanem, P. Rosso, and F. Rangel, "An Emotional Analysis of False Information in Social Media and News Articles," ACM Trans. Internet Technol., vol. 20, no. 2, May 2020, doi: 10.1145/3381750.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY