iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55564

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 11 Issue VIII Aug 2023- Available at www.ijraset.com

Recognition of Phony Reviews in E-commerce Platform

Raghavendra Babu T M [1], Harshith K S[2]
Department of computer science and Engineering, Department of Information science and Engineering, PES College of Engineering, Mandya, India

Abstract: ***The business and commerce of today are greatly influenced by online reviews. The majority of how consumers choose which online products to buy is based on user reviews. As a result, opportunistic people or organizations try to slant product reviews in order to serve their own agendas. Now days, fake online reviews are more common, people simply post phony reviews to market their products, to cheat customers etc. In E-commerce application fake reviews are more common and we need a system to predict these fake reviews. Proposed system uses machine learning algorithm to identify the fake reviews in an E-commerce platform. We build a new E-commerce application and customer registers themselves, browse and purchase the products. For the purchased products customers post reviews, and our system will predict the fake reviews using an algorithm.***
Keywords: E-commerce, Phony reviews, Customer behavior, purchasing patterns, Prediction of reviews.

I. INTRODUCTION

In the previous two decades, technology has advanced quickly. New and advanced technologies are constantly replacing obsolete ones. People can now do their work effectively thanks to new technologies. The online market is one such example of this technological advancement. Using internet, we may purchase and make reservations on every commodity. Nearly every one of us reads reviews before deciding to buy a product or service. Online reviews are now a valuable source of reputation for businesses. Additionally, they play a significant role in product and service promotion together with advertising. Phony internet reviews are a growing concern as online marketplaces grow in popularity. For the purpose of promoting their own items, someone may write deceptive evaluations that are harmful to the real users. In addition, competitive businesses may attempt to harm one another's reputations by posting fabricated testimonials.

Numerous methods for identifying these phony reviews have been studied by researchers. While some techniques are based on the user behaviour who posts the reviews, others are based on the content of the reviews. The IP address, the number of postings, and other user behaviour-related data are the focus of the user behaviour-based method, whereas the review's text is the subject of the content-based study. Most of the mentioned techniques rely on supervised classification models. Semi-supervised models have only been employed by a small minority of researchers. In order to address the reviews' unreliable labelling, semi-supervised approaches are being deployed. In our work, we develop several semi-supervised and supervised classification methods for identifying phony reviews. In order to enhance classification performance, we used Naive Bayes classifiers in our research. We have mainly concentrated on the review-based techniques on the content. Word frequency, emotion polarity, and review length were employed as features. Algorithms for Machine Learning are typically based on Mathematics and Statistics. The key benefit of machine learning over traditional software is that there isn't any written code telling the system how to decide between two objects because it's difficult to handle every circumstance on an item. Such circumstances benefit from machine learning. It mostly makes use of the provided data to make wise decisions, make forecasts for the future, or find abnormalities. It is used in many applications today, including advertising, predictive analysis, search engine results, virtual assistants, object identification, autonomous vehicles, and more. Supervised Learning, Unsupervised Learning, and Reinforcement Learning are the three subfields that make up the term "Machine Learning techniques".

II. RELATED WORK

As per the authors Arush Agarwal and Akhil Dixit in their work Hate Comments Detection: An Ensemble Learning Approach, phony reviews are described as a form of misleading journalism and claims that are made to deceive and mislead consumers. Additionally, the legitimacy of social media sites, where this news is primarily spread, is in jeopardy. The limitation of their work is that they use small dataset yielding less accurate results. Further they use SVM, Regression techniques which generate graphical outputs, not suitable for real time application [1].

The authors of a smart system for false news detections using machine learning (ML), Anjali Jain, Avinash Shakya, Harsh Khatter, and Amit Kumar Gupta, claim to have established a model and the approach for fake news detection. They attempted to aggregate the news with the use of machine learning and natural language processing in order to eventually use Support Vector Machine to decide if the news is true or false. The Limitations of their work are that they use SVM, Regression techniques which generate graphical outputs, not suitable for real time application, works fine for small datasets, takes more time prediction, less efficient and applicable for only news [2]. Consumers are increasingly rating, reviewing, and researching things online, according to authors Claire Cardie and Jeffrey T Hancock in Finding deceptive opinion spam by any stretch of the imagination. As a result, websites that host customer reviews have become targets for opinion spam. While the majority of recent research has focused on manually identifying instances of opinion spam, they looked at misleading opinion spam, which is composed of phony opinions that are intentionally manufactured to look real. Combining work from computational linguistics and psychology, they constructed and evaluated three methods for identifying false opinion spam, ultimately creating a classifier that is over 90% accurate on their benchmark dataset for false opinion spam. They also made various theoretical contributions based on feature analysis of their learnt models, such as demonstrating a link between false ideas and innovative writing [3]. Authors S. Feng, R. Banerjee, and Y. Choi introspected that previous computational fraud detection research has mostly relied on shallow lexico-syntactic patterns in their work Syntactic stylometry for deception detection. By examining syntactic stylometry for deception detection, the authors contributed a distinctive viewpoint to the subject. They demonstrated across four different datasets, including essays and product reviews, that features obtained from Context Free Grammar (CFG) parse trees considerably improve identification accuracy over multiple baselines based only on shallow lexico-syntactic characteristics [4].

The authors E.P. Lim, V.A. Nguyen, N. Jindal, B. Liu and H.W. Lauwin Detecting product review spammers using rating behaviors aim to recognize people that generate spam review so review spammers in their paper. They found numerous common patterns among review spammers and created a model to detect the spammers. They are particularly interested in modeling the following behaviors. Tomaximize their influence, spammers may first target certain items or product groupings. Second, they have a tendency to differ from other reviewers in their product ratings. The degree of spam for each reviewer on an Amazon review dataset was determined using scoring methods that were proposed and tested. They then employ a web-based spammer evaluation tool created specifically for user evaluation trials to pick a selection of highly suspect reviewers for further investigation by their user evaluators. The findings demonstrate that our suggested ranking and supervised approaches beat previous baseline methods based solely on helpfulness votes in terms of identifying spammers. It demonstrates, at long last, that compared to unhelpful reviewers, discovered spammers have a greater impact on ratings [5]. In the work of Revisiting semi-supervised learning for online misleading review identification, authors J. K. Rout, A. Dalmia, and K. R. Choo claim that more customers are using online reviews to guide their service decisions and that these reviews have a financial influence on companies' bottom lines. It should come as no surprise that shady individuals or groups have tried to take advantage of or manipulate online opinion reviews (such as spam reviews) in order to profit or achieve other ends, and that identifying dishonest and phony opinion reviews is a subject of current study interest. Before proving its usefulness with a data set of hotel reviews, the authors describe how semi-supervised learning techniques can be used to detect spam reviews [6].

A. Gaps Analyzed

1) Most of the works predict fake reviews based on the training datasets using ML techniques, but are not implemented in E-commerce platform.

2) In addition, they put forth the notion of predicting bogus reviews, but in the system, we propose, we develop this idea into a real-time application helpful to customers and commercial entities.

3) Many of the works have used algorithms such as *SVM*, regression and other algorithms which generates less accurate results and less parameters used for prediction, but in out proposed system we use efficient algorithms and around 10 parameters used.

4) We use efficient algorithms such as "*Bayesian algorithm*", or "*Decision Tree algorithm*" or "*KNN algorithm*" to get good results.

5) In contrast to all of the earlier research, the suggested method uses enormous data sets for processing.

6) Some of the works used tools and libraries for algorithms that had already been developed rather than creating their own algorithms from scratch. But under the proposed system, the algorithm is programmed, requiring us to develop our own rationale for it and validate the results.

III. EXISITING WORK

A lot of research works are done on this phony reviews' detection using certain algorithms. Most of the works use ready tools and ready packages for detection, tools such as "WEKA", "R-Tool", "RAPID MINER" etc. None of the existing works implemented this concept in real time. By real time we mean, detection of phony reviews on E-commerce platform as soon as the customer posts some reviews about a particular product. Although few works use ready tools and libraries, results obtained are less accurate and takes more time for data processing. In existing systems, content-based methods are used which focus on what is the content of the review such as text of the review or what is told in it. Under this category, three techniques are used namely Genre Identification, Detection of Psycholinguistic Deception and Text categorization. In all of these previous works, assuring the quality of the datasets and the reviews is difficult which results in poor predictions of the reviews.

IV. PROPOSED WORK

Instead of concentrating completely on the content and texts of the reviews, we mainly focus on the behavior and the information pertaining to the reviewer to build our system model. People who post intentional phony reviews have significantly different behavior than the normal user. Due to the lack of proper dataset available in order to train the system, we are building our own e-commerce platform. The system will use information such as the purchases of the customers and we can get to know about the location and interests of the customers when they try to register on our e-commerce platform. This will help our classifier to identify some of the phony reviews. Finally, the system will also consider situations where customers constantly change their user's name and give reviews to various products. Parameters like the time and frequency of posts, percentage of positive and negative reviews combined with the types of words used, sentence formation styles of the reviewer will also be considered by the proposed classifiers. This way we have produced a better classifier which gives us a better result compared to the existing works.

V. IMPLEMENTATION

A. Architectural Design Of Our E-Commerce Platform

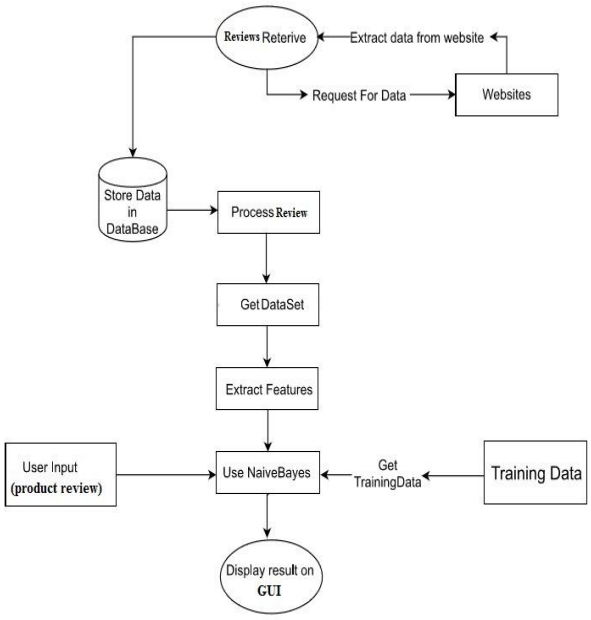


Fig 1: Architectural design of E-commerce platform

In the Fig 1, we see the working design of our system. Here, the customer interacts with the E-commerce platform and get the data pertaining to the customer. The testing data is used by the Naïve Bayes algorithm for the classification and hence the prediction. We primarily use Naïve bayes algorithm as a base to construct our proposed system and then go on adding additional features up on it. A group of classification algorithms built on the Bayes Theorem are known as Naive Bayes classifiers Instead of being a single algorithm, it is a family of algorithms, and they are all predicated on the notion that every pair of features being classified is unrelated to every other pair.

The steps on how it actually works are as shown below:

1) *Step 1:* Scan the dataset (storage servers).
2) *Step 2:* Calculate the probability of each attribute values [n, n_c,m,p].
3) *Step 3:* We apply the formulae

$P(\text{attributuevalue}(a_i)/\text{subjectvalue}(v_j))=(n\_c+mp)/(n+m)$

Where:

- n = the number of training examples for which $v=v_j$
- n_c=number of examples for which $v=v_j$ and $a=a_i$
- p=a priori estimate for $P(a_i,v_j)$
- m = the equivalent sample size

4) *Step 4:* Multiply the results of each attribute p and utilize the total values for each class when calculating the probability for each class.
5) *Step5:* Compare the values and classify the attribute values to one of the predefined sets of class.

B. Classification Rules

In essence, classification is used to assign each piece of information in a batch of data to one of the established classes or groups. Mathematical methods are used in classification procedures to solve problems. For instance, a company's employee statuses (leaves or stay) In our work, we process historical data and make predictions using either "Naive Bayes," "KNN," or "Decision Tree" classifiers. These particular algorithms are the most effective and process data more quickly. These algorithms operate correctly with n different parameters.

VI. EXPERIMENTAL RESULT

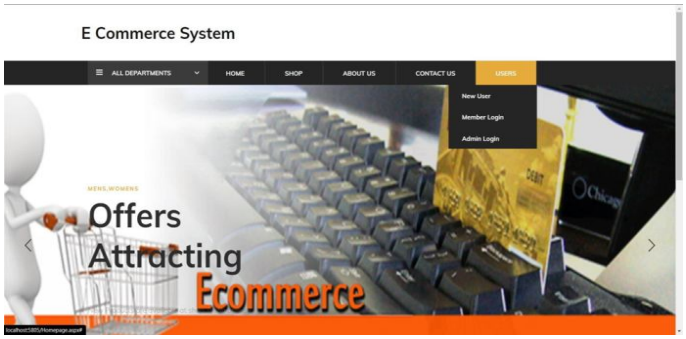


Fig 2: The Homepage of our E-Commerce platform

The above Fig 2 shows the home page of our E-commerce platform which is viewed by the visitor.

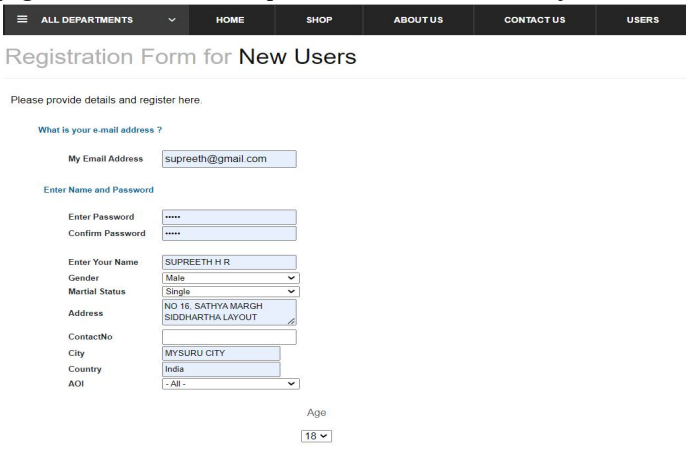


Fig 3: Registration Page for the new user

The above Fig 3 depicts the registration of the new customer in order to capture the details of the customer.

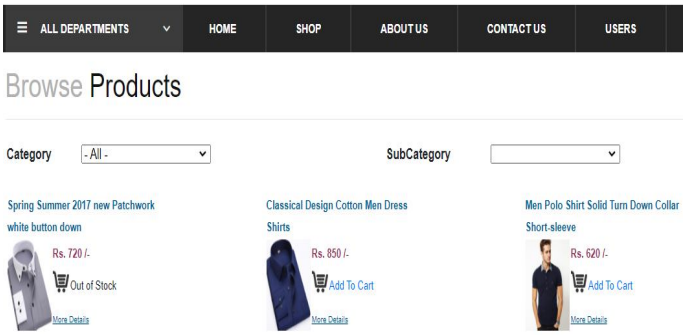


Fig 4: The Product available for the members to buy

The above Fig 4 depicts the products that the registered members can buy in which the products are displayed category wise.

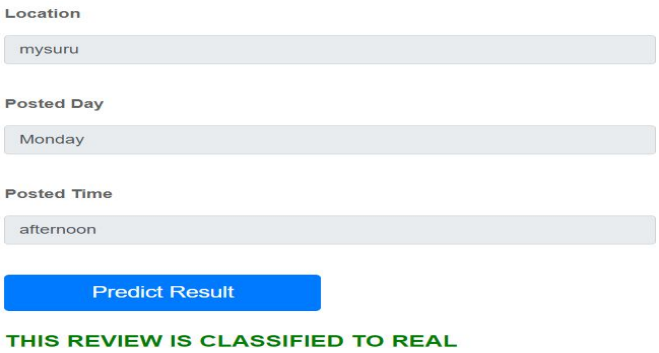


Fig 5: The Classification of member reviews as fake or real

The above Fig 5 shows the prediction of fake reviews given by the customer on the bought products.

Constraint	Naive Bayes Algorithm
Accuracy	96.25%
Time (milli secs)	492
Correctly Classified	96.25%
InCorrectly Classified	3.75%

Fig 6: The accuracy of Naïve Bayes algorithm

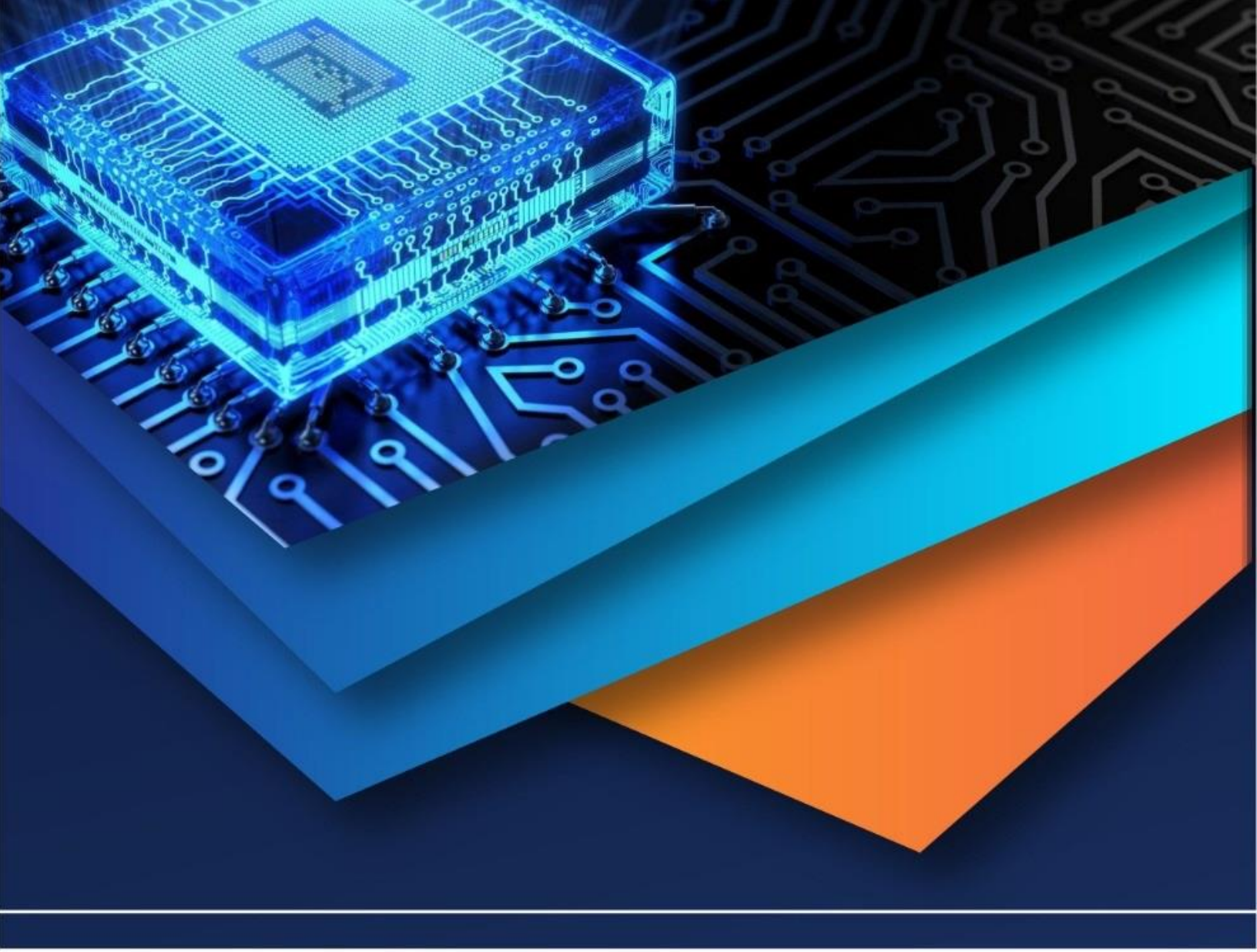
The above Fig 6 displays the accuracy of the Naïve Bayes algorithm with the time taken to classify and the number of reviews classified correctly. After running the naïve bayes algorithm on the admin side of the E-commerce platform, the testing data is classified and the results are then compared with the actual dataset stored in the database of our work. The algorithm takes in the various aforementioned parameters which it then calculates the probability for each parameter. Based on the probability, the algorithm then classifies the reviews as 0 for real and 1 for fake.

VII. CONCLUSION

The goal of every firm is to maximize profits. Many features are available in today's e-commerce platforms, such as "flipkart.com," "amazon.com," "myntra.com," and others, to aid customers in their buying. Identifying the fake product reviews is a challenging task in the current e commerce sector. System detects the fake product review by using data science algorithms. Our Proposed system detects fake reviews in an efficient way and generates more accurate results.

REFERENCES

[1] Arush Agarwal, Akhil Dixit.T, "Hate Comments Detection: An Ensemble Learning Approach, 2019.
[2] Anjali Jain, Avinash Shakya, Harsh Khatter, Amit Kumar Gupta, A smart system for fake news detections using ML, 2015.
[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination", in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT),vol.1,pp.309–319,Association for Computational Linguistics, Portland, Ore, USA,June2011.
[4] S.Feng,R.Banerjee,andY.Choi,"Syntacticstylometryfordeceptiondetection", in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Vol. 2, 2012.
[5] E. P. Lim, V.A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proceedings of the 19th ACM International Conference on information and Knowledge Management(CIKM), 2010.
[6] J. K. Rout, A. Dalmia, and K.-K. R. Choo, "Revisiting semi-supervised learning for online deceptive review detection", IEEE Access, Vol. 5, pp. 1319–1327, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY