FAKE REVIEW DETECTION IN E-COMMERCE : A MACHINE LEARNING APPROACH USING NATURAL LANGUAGE PROCESSING

Nandhini R[1], Krishnaveni A[2], Ragunathan S[3]

ABSTRACT

Techniques using Natural Language Processing (NLP) to identify and remove bogus reviews are highly useful in the e-commerce sector. In this work, we employ two different machine learning models - Naive Bayes and XGBoost - to classify whether a given dataset is authentic or not. Fake reviews undermine user trust, making this a critical challenge for e-commerce companies to address. By building this model, it can help the owners of the websites and the application developers identify these fake reviews and therefore can preserve the consumer's confidence in online shopping. In this paper proposed Naive Bayes and XGBoost classifiers, the model very efficiently gives out the results of the probabilities of the likelihood of the review to be fake or not. Although trained, only Amazon and Yelp datasets achieved an impressively high accuracy score using XGBoost. Its scalability suggests that using this model on larger datasets could further enhance its accuracy and adaptability in identifying fake reviews.

Keywords: Sentiment analysis, opinion mining, text mining, fraudulent review detection, machine learning, NLP, Naive Bayes, Random Forest.

I. INTRODUCTION

The impact of online reviews on consumers' purchasing decisions has been greatly increased by the quick growth of e-commerce. Since customers cannot physically scrutinize products prior to purchase, a growing number of customers rely on views and experiences provided by other users [1]. However, the growing popularity of false or misleading reviews often leads the customers astray in making poor buying decisions. It is impossible for the customers to individually check and authenticate each review as it tends to be very time-consuming with the ever-growing volume of online reviews. To address the challenge, ML techniques combined with Natural Language Processing would hopefully scan review content to determine whether or not they are authentic. Fake reviews often exhibit recognizable language patterns, such as exaggerated praise or repeated expressions like "awesome," "fantastic," or "so good." However, identifying attempts to artificially elevate a product's rating is challenging, as these patterns extend far beyond simple sentiment analysis. Some of the subtle factors involved are sarcasm, delivery issues, or even phrases for expressing dissatisfaction masked under seemingly pleasant words, making classification difficult, hence requiring complex techniques in order to detect them effectively [2].

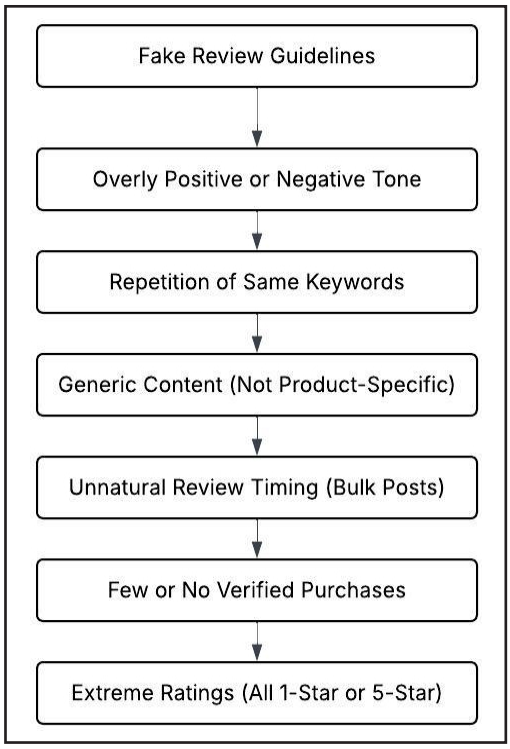

Fig 1 : Fake review guidelines from Times of India

The techniques involved in NLP are essential in identifying such subtleties and improving the accuracy rate for fake review detection systems. Once preprocessed, delete reviews that are old or irrelevant. The goal is to create an E-commerce environment that consumers could trust in, thereby making the opinion read on the website trustworthy and so also the products purchased from it. This is extremely relevant for big

Department of Computer Science[1],
Karpagam Academy of Higher Education, Coimbatore, India[1]
nandhini.raghumurugan@kahedu.edu.in[1]

Department of Computer Science[2]
Karpagam Academy of Higher Education, Coimbatore, India[2]
krishnaveni.arumugam@kahedu.edu.in[2]

Department of Computer Science[3]
AVS College of Arts & Science (Autonomous), Salem, India[3]
ragu34salem@gmail.com[3]

* Corresponding Author

E-commerce websites like Amazon and Flipkart where large numbers of users mean a greater necessity for the integrity of user reviews among all parties concerned. In addition to E-commerce, fake reviews and misinformation abound in many industries such as travel (e.g., TripAdvisor) and food delivery (e.g., Swiggy, Zomato) and social media platforms like Twitter and Facebook using sentiment analysis to flag fake content [3].

Similarly, sophisticated measures to protect consumers against misleading reviews are adopted through these E-commerce platforms. As previous studies have shown, manual annotation of reviews is inefficient; it takes researchers some several weeks to label small datasets. It also means that scalable, automated solutions that will involve supervised machine learning models are strongly required [4].

This paper presents a comparative study of two kinds of machine learning models like Naive Bayes and XGBoost. These machine learning models are used to detect fake reviews with the use of a subset of both Amazon and Yelp datasets. Specifically, XGBoost has achieved promising accuracy at 89.81% while outperforming Naive Bayes in the task of fake review detection. The results underscore thus the importance of selecting the right features and appropriate model training to improve the accuracy of detection. Therefore, the study presents a complete solution for its application in the removal of fake reviews, and findings present for real-time implementation in E-commerce systems.

II. LITERATURE REVIEW

In [5], Barbosa, L., & Feng, J. (2010) made significant contributions to sentiment analysis on social media by addressing two key challenges in Twitter data; bias from self-selected users and noise from informal content. The authors developed a two-step approach combining meta-classification (using multiple existing sentiment tools) with Twitter-specific feature selection, introduced innovative elements such as syntax features from POS tagging and Twitter-specific features like hashtags and retweets. This work pioneered robust methods for social media text analysis and emphasized the importance of domain-specific feature engineering, though it had limitations such as dependency on existing sentiment tools and potential scalability issues. Despite these constraints, the paper significantly influenced subsequent social media sentiment research by establishing the need for specialized approaches to handle the unique characteristics of social media text [5].

In [6], Patel, D, et.al (2018) presented a significant contribution to opinion mining and e-commerce fraud prevention. The study developed a comprehensive framework combining text preprocessing, feature extraction, and classification algorithms to identify fraudulent online reviews. The authors identified key indicators of fake reviews, including extreme sentiment, suspicious temporal patterns, reviewer behavior, and linguistic markers. The methodology employed various machine learning classifiers and analyzes different feature sets effectiveness in fraud detection. While the research is limited by its dataset size and the potential for evolving deception techniques, its significance lies in practical applications for e-commerce platforms and its contribution to online trust systems. The paper effectively built upon existing sentiment analysis work while specifically addressing the growing challenge of detecting increasingly sophisticated fake reviews, representing an important advancement in applying opinion mining to real-world problems [6].

In [7], Wahyuni, E. D., & Djunaidy, A. (2016) presented a significant advancement in fake review detection through a modified iterative computation framework. The research uniquely combines reviewer behavior analysis with content-based features, offering enhanced computational efficiency and accuracy in identifying fraudulent e-commerce reviews. This methodology employed a multi-step process involving initial content-based scoring followed by iterative refinement of trustworthiness scores, integrated with reviewer credibility metrics. The study demonstrates notable advantages in reduced computational complexity and improved scalability for large-scale review systems, though it faces limitations such as dependency on initial parameter settings and potential sensitivity to evolving review patterns. While the research makes a valuable contribution to the field of opinion spam detection by balancing computational efficiency with accuracy, the authors acknowledged the ongoing challenge of adapting to sophisticated deception techniques. The paper's significance lies in its practical applicability to real-world e-commerce systems, offering a methodological advancement that builds upon previous iterative computation approaches while addressing key efficiency and accuracy concerns [7].

In [8], Ren, Y., & Ji, D. (2017) published in Information Sciences, presents an empirical investigation into the application of neural networks for detecting deceptive opinion spam. The research addresses the growing challenge of fake reviews in e-commerce platforms by leveraging deep learning techniques. The authors likely explored various neural network architectures, comparing their effectiveness in identifying fraudulent opinions against traditional machine learning methods. The study presumably evaluated different feature extraction approaches and their impact on detection accuracy, while also considering the computational

requirements and scalability of neural network-based solutions. This work contributed significantly to the field of opinion spam detection by providing empirical evidence of neural networks' capability to learn complex patterns in deceptive text, potentially offering more robust and adaptive solutions compared to conventional approaches. Findings of our research have important implications for e-commerce platforms and content moderation systems, advancing the state-of-the-art in automated deception detection [8].

In [9], Joseph et.al (2019) analyzes the important algorithms based on factors such as computational efficiency, scalability, and accuracy, while addressing current limitations and future research directions. Their paper's significance lies in its thorough examination of how data mining algorithms contribute to intelligent computing systems, serving as a valuable resource for researchers and practitioners navigating the challenges of extracting meaningful insights from increasingly complex and voluminous data environments [9].

In [10], the researchers effectively identified that fraudulent reviews while protecting data holders' privacy, Jianto et al. (2025) suggested a federated learning framework that is sensitive to data quality. In particular, they added a module for evaluating the quality of the input when training federated learning. This module uses a variety of criteria, such as user behaviour, text completeness, and annotation correctness, to quantify the quality of each client's data In [11], Anas, S & Kumari, S (2021) introduced a complex model that has been trained on millions of reviews is needed to combat these scammers. The models in this work are trained using the "amazon Yelp dataset," which is a very small dataset that can be scaled to achieve high accuracy and flexibility.

In [6], Patel D et.al (2018) suggested approach uses phase-wise processing to automatically classify user reviews into "suspicious," "clear," and "hazy" categories. Elements are iteratively eliminated into suspicious or clear categories in the hazy category in [12], Punde et.al (2019) utilized machine learning, they developed a system that detects and removes all fraudulent reviews. Additionally, they exclude reviews that are flooded by a marketing firm to raise a product's ratings

In [13], Saumya and Singh looked into each of these aspects for a questionable review list, which only included reviews that had peer user comments. The F1-score for the suggested system was 91%. Since they suggested a technique may be used independently to clean up product review datasets, it can be a huge help in spam detection systems. Heydari A et.al (2016) suggested approach might be a great help for online spam filtering systems and could be used as a stand-alone system to clean up product review datasets for data mining and knowledge discovery tasks. Their approach can help these systems in terms of high accuracy and time efficiency [14]. Rintu Augustine and Krishnaveni Arumugam (2025) examined ways to counteract blockchain-related threats, such as BGS, GNNs, selfish mining, and double spending. They concluded that while GNNs are quite good at identifying fraudulent transactions, they still have issues with scalability, dynamic transaction patterns, and over-smoothing [15]. Predicting a web user's next page from server log data is a central task in web usage mining, with applications in prefetching, personalization, site re-structuring, and recommendation. At its core the problem requires converting raw server logs into meaningful sessions and sequences, then modeling those sequences to estimate the most likely subsequent request. [16]

III. PROPOSED SYSTEM

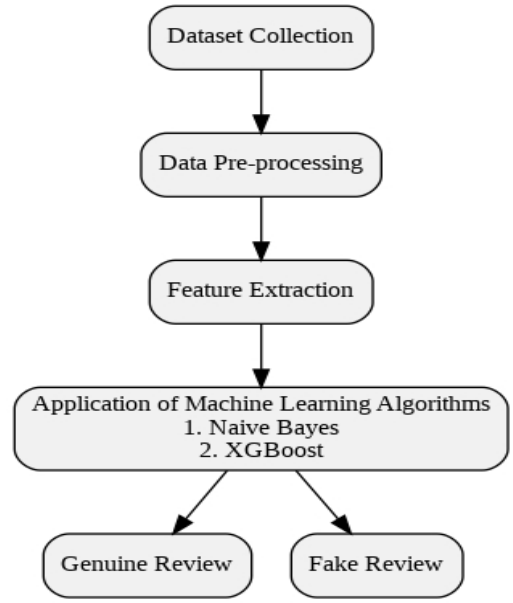

Fig 2 : Block Diagram of Overall Methodology of Proposed System

The above Fig 2 describes the following components :

1) Gathering Datasets
2) Preparing Data
3) Extraction of Features
4) Using Machine Learning Techniques
5) Genuine Review
6) Fake Review

Flow of the Proposed System

As depicted in the Fig 2, this sentiment extraction technique will be performed by the following steps:

A. Dataset Collection:

This study uses the "Amazon Academic Review" dataset that contains attributes such as reviews, useful votes, ratings, user IDs, and many other features. Relevant parameters are extracted during feature engineering so that models could be accordingly trained and evaluated in an effective way. Dataset here is comprised of a mix of genuine and fake reviews; therefore, results can be best assessed for the actual performance of the model. It uses the Yelp dataset that was made public as part of an academic challenge. The number of businesses in this dataset is 11,537, check-in sets 8,282, users 43,873, and reviews 229,907; it is accessible at www.yelp.com/dataset. The diversity of a great many reviews and attributes makes this dataset particularly challenging to face, thus making it apt for training machine learning algorithms in detecting fake reviews.

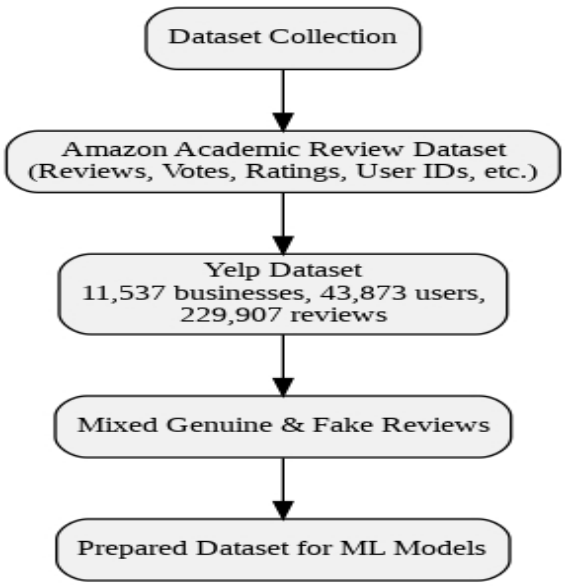

Fig 3: Dataset Collection

B. Data Pre-processing

The first preprocessing step, which is quite crucial for the analysis of any dataset, consists of data cleaning in which unnecessary attributes along with punctuation, stop words, missing values, and redundant words are removed. The stage ensures proper preparation of the dataset that will train for enhancing the overall effectiveness as well as accuracy of the model.

Table -1: Data Pre-processing

Column	Missing Values
category	0
rating	0
label	0
text_	0

C. Feature Extraction :

The methods of eliminating unwanted information from the dataset are termed as data cleaning. This is the most crucial step involved in the process, which identifies gaps and further explains various relations between the attributes (columns) that may draw valid conclusions about whether the review is authentic or not. From the perspective of fake review detection, some libraries taken from NLTK package were used to build a bag of words that is suitable as a corpus for analysis. The most important critical functions utilized include term frequency, tokenization, and stop word removal. Stop words such as "is," "then," "to," and "why" are grouped together because they are redundant and contribute little value to the feature engineering process. Therefore, in the context of detecting fake reviews, recognizing these words provides no additional useful information. The frequency count calculates the occurrence of each word in the reviews. In this sense, it helps identify spamming pattern presentation. With that, the frequencies analyzed will be able to recognize more fraudulent reviews, making the E-commerce much more reliable.

D. Data Sampling

Due to the vast number of reviews under consideration in the dataset, it employs data sampling before presenting the data to the classifier. Data sampling reduces the computational burden borne by the classifier to digest the data in portions for easier processing. In sampling, the labels may vary that distinguish between the authentic and spurious reviews. Once the column name "short_str" is assigned, appropriate columns are connected and return the resulting Data Frame. This approach ensures that the classifier is trained effectively by having a balanced combination of authentic and fraudulent reviews to enhance the accuracy rate of the detection model for fake reviews in E-commerce.

E. Methodology:

Machine Learning algorithms have been used in this proposed system, such as Naïve Bayes and XGBoost.

1. Naive Bayes :

This "naive" assumption simplifies the computation so that the classifier works efficiently even when the dimensionality of input data is high. Naive Bayes is based on Bayes' theorem, which calculates the probability of a review being spam or genuine based on the features extracted from the review text. This technique is very useful for real-time prediction and can efficiently predict the probability of multiple classes of target attributes such as authentic or

fraudulent reviews. In the case of detecting fake reviews, the first step involves determining the prior probabilities of each class in the dataset often referred to as class probabilities. From this, conditional probabilities are calculated, which represent the possibility of each attribute occurring given a particular class. From these probabilities, the Naive Bayes classifier can now effectively detect and flag fake reviews, thereby increasing the credibility of the E-commerce websites.

Naive Bayes (NB) was applied in the following way:

1. Importing library Multinomial NB from sklearn. naive_bayes
2. Now we create Naive Bayes Classifier object
3. We fit our model lastly with our data

2. XGBoost :

XGBoost is a highly optimized version of the gradient boosting algorithm. It has developed especially in order to boost efficiency and to make it much faster than any other general approach, so that its overall performance on models is excellent. In the context of E-commerce's review-fake-detection, the superiority of XGBoost in producing results is well known than any other machine learning algorithm. XGBoost is an open-source library built under the auspices of Distributed Machine Learning Community, where researchers and developers can tap into it for a wide range of applications. It makes use of parallel tree boosting or Gradient Boosted Decision Trees GBDT or GBM. It deals very well when it comes to the handling of large data and dealing with complex relationships that happen to be between the different data, and therefore, it is one of the best models for detecting real or fake reviews. The ability of XGBoost to perform well in classification tasks makes it greatly beneficial for making online platforms more reliable and boosting the consumer's trust in E-commerce.

We have applied XGBoost in our model as:

1. Importing library XGBClassifier from XGBoost
2. Now we create XGBClassifier object
3. In the last we fit our data

IV. EXPERIMENTAL RESULTS

The data table presents a comparative analysis of three machine learning models across three iterations or nodes. The Proposed XGBoost model consistently outperforms the other two models, starting with a strong 79.007% accuracy in the first iteration, improving to 82.80% in the second, and reaching an impressive 89.81% in the third iteration. The Ex1 Base Model shows steady improvement throughout the iterations, beginning at a modest 59.00%, increasing to 68.50% in the second iteration, and achieving 75.70% in the final iteration. In contrast, the Ex2 Naive Bayes model exhibits more variable performance: it starts at 70.10%, peaks at 79.34% in the second iteration, but then experiences a significant drop to 67.60% in the third iteration. This detailed breakdown of performance metrics clearly demonstrates the superiority of the Proposed XGBoost approach, which not only maintains the highest accuracy throughout but also shows consistent improvement across all iterations, unlike the more erratic performance of the Naive Bayes model or the slower progression of the Base Model.

Table -2: Proposed XG Boost Algorithm

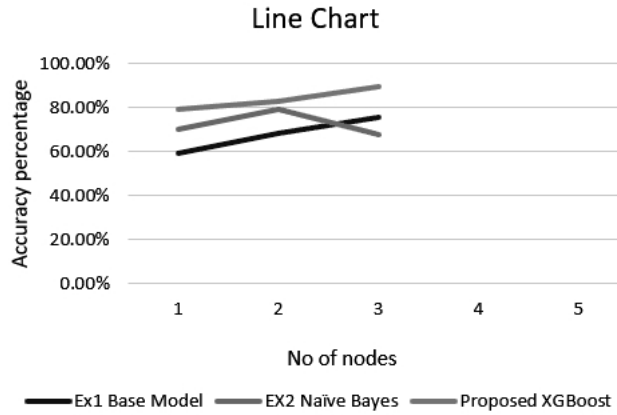

Fig 5: Comparison Chart

Figure 5 The line chart illustrates a performance comparison between three different models: Ex1 Base Model, Ex2 Naive Bayes, and a Proposed XGBoost approach, measured across varying numbers of nodes. The accuracy percentage, ranging from 0% to 100%, is plotted against the number of nodes, which extends from 1 to 5 on the x-axis. The Proposed XGBoost model, represented by the gray line, demonstrates superior performance, starting at approximately 75% accuracy and steadily climbing to about 85% by the third node. In contrast, the Ex2 Naive Bayes model (orange line) initially begins around 70% accuracy, shows a slight improvement, but then experiences a decline after the second node. The Ex1 Base Model, shown in blue, starts with the lowest accuracy at roughly 60%, sees an increase to about 75% at the second node, but subsequently drops in performance. Notably, the data visualization appears incomplete, only showing results up to the third or fourth node despite the x-axis extending to five nodes. Overall, the graph effectively illustrates the Proposed XGBoost model's consistently better performance and upward trajectory compared to the other two models, suggesting it may be the most promising approach among the three tested methods.

Table - 2 : Proposed XG Boost Algorithm

Ex1 Base Model	Ex2 Naive Bayes	Proposed XGBoost
59.00%	70.10%	79.007%
68.50%	79.34%	82.80%
75.70%	67.60%	89.81%

V. CONCLUSION AND FUTURE SCOPE

The results clearly point out that XGBoost performed better than the Naïve Bayes algorithm, but at the same time gives evidence that this model allows correctly identifying fraudulent reviews. Based on this experiment, there is critical evidence that the choice of the proper algorithms to detect and remove deceptive reviews properly is necessary, although the task is challenging in itself. Such knowledge could greatly improve the quality of an online sale process and, as a result, increase consumer trust. The rapid growth of E-commerce requires effective mechanisms for detecting fake reviews so that there can be a fair marketplace. Future research directions could focus towards hybrid models with the fusion of different algorithms that may lead to enhanced accuracy in detection.

REFERENCES

[1] Chen T, Samaranayake P, Cen X, Qi M and Lan Y-C, "The Impact of Online Reviews on Consumers Purchasing Decisions: Evidence from an Eye-Tracking Study", Front. Psychol. 13:865702. (2022) doi: 10. 3389/fpsyg. 2022.865702

[2] Shuvo Kumar Mallik, Imran Uddin, Farzana Akter, A. S. M.Shafin Rahman, M Abeedur Rahman, "Evaluating the influence of customer reviews and consumer trust on online purchase behavior", WJARR, (2025), 25(01), 423-432 //doi.org/10.30574/wjarr.2025.25.1.0015

[3] Shenil Polpolage, "Fake Review Detection in Yelp Restaurant Reviews via Natural Language Processing", Research Square (2025), https://doi.org/10.21203/rs.3.rs-6305783/v1

[4] Mohammed Ennaouri, Ahmed Zellou, "Machine Learning Approaches for Fake Reviews Detection: A Systematic Literature Review", Journal of Web Engineering, Vol. 22 5, 821–848. doi: 10.13052/ jwe 1540-9589. 2254 (2023)

[5] Barbosa, L., & Feng, J, "Robust sentiment detection on twitter from biased and noisy data In Coling", Posters (pp. 36-44), (2010)

[6] Patel, D., Kapoor, A, Sonawane, S, "Fake review detection using opinion mining", International Research Journal of Engineering and Technology (IRJET), 5(1), 192-201. (2018)

[7] Wahyuni, E. D., Djunaidy, A, "Fake review detection from a product review using modified method of iterative computation framework", In MATEC web of conferences (Vol. 58, p. 03003). EDP Sciences, (2016)

[8] Ren, Y., Ji, D, "Neural networks for deceptive opinion spam detection: An empirical study", Information Sciences, 385, 213-224, (2017)

[9] Joseph, S. I. T., Thanakumar, I. "Survey of data mining algorithms for intelligent computing system", Journal of trends in Computer Science and Smart technology (TCSST), 1(01), 14-24. (2019)

[10] Jiantao Xu, Chen Zhang, Liu Jin, Chunhua Su, "Data Quality-Aware Federated Learning for Fake Review Detection", 2025 7th International Conference on Software Engineering and Computer Science (CSECS), pp.1-6, (2025)

[11] Anas, S. M., & Kumari, S. "Opinion mining based fake product review monitoring and removal system", 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 985-988). IEEE, (2021)

[12] Punde, A., Ramteke, S., Shinde, S., Kolte, S, "Fake product review monitoring & removal and sentiment analysis of genuine reviews", International Journal of Engineering and Management Research (IJEMR), 9(2), 107-110. (2019)

[13] Saumya, S., Singh, J. P, "Detection of spam reviews: a sentiment analysis approach", CSI Transactions on ICT, 6(2), 137-148, (2018)

[14] Heydari A, Tavakoli M, Salim N, "Detection of fake opinions using time series", Expert Syst Appl 58:83–92, (2016)

[15] Rintu Augustine, A.Krishnaveni, "Enhancing Blockchain Security: A Comprehensive Study on Fraud Detection and Prevention", Karpagam JCS, 20 (03), 165-170 May-June (2025)

[16] Jothish Chembath, E.J.Thomson Fredrik, An Empirical Analysis of Algorithms to Predict Next Web Page Using Web Log Data, International Journal of Applied Engineering Research, Vol.12, No.16, 2017