International Journal of Innovative Research in Engineering
Volume 3, Issue 6 (November-December 2022), PP: 53-55.
www.theijire.com **ISSN No: 2582-8746**

A Survey on Fake Customer Review Detection System

Abhishek Kumar Roy[1], Devsharan Singh[2], Imran Raeeni[3], Izmamul Ansari[4], Vijendra Pratap Singh[5]
[1,2,3,4] B. Tech Student, Computer Science and Engineering, Institute of Technology and Management, Gorakhpur, U.P., India.
[5] Assistant Professor, Computer Science and Engineering, Institute of Technology and Management, Gorakhpur,U.P., India.

How to cite this paper: Abhishek Kumar Roy[1], Devsharan Singh[2], Imran Raeeni[3], Izmamul Ansari[4], Vijendra Pratap Singh[5], "A Survey on Fake Customer Review Detection System", IJIRE-V3I06-53-55.

Abstract: *As we know, now a day's online shopping become a daily activity for humans. Before going to buy any product in e-commerce business organizations. Reviews are the one of the important ways to check reliability of a product. Customer will check reviews posted by other customers to buy a product. If a customer bought a product by seeing fake review, if the product is really good no problem otherwise a product loses its reliability. We are here to perform sentiment analysis on restaurant reviews to find number of correct and number of wrong predictions made by the classifier which is further helpful to classify reviews into real or fake.The classifiers used in our project are Natural Language Processing, Support vector Machine (SVM), and Naıve Bayes. The measured results of our experiments show that the SVM algorithm outperforms other algorithms, and that it reaches the highest accuracy not only in text classification, but also in detecting fake reviews.*

Key Word: *Supervised Machine Learning Techniques, Support Vector Machine, Natural Language Processing and Naıve Bayes.*
.

I. INTRODUCTION

As more people can access services and goods online, online shopping is gradually increasing. More individuals are responding to what sellers are saying about their businesses, which has irritated some people who then mislead others by spreading untrue information in an effort to promote or harm a particular commodity or service's reputation. These individuals are referred to as perception spammers, and the phony testimonials they leave are regarded as fake remarks. Although customer reviews may be useful, placing naive trust in them could be dangerous for both consumers and sellers. Before making any online purchase, many shoppers read research. Additionally, the remarks might be deceptive in order to gain additional advantage or profit, therefore any purchasing decision based on online comments should be carefully considered.

Opinion spamming comes in a few distinct forms. One type is endorsing certain products with the purpose to encourage endorsing products with misleading or unfavorable evaluations to harm their reputation. The second category consists of product-neutral advertisements. There has been a lot of research in the area of sentiment analysis, and models have been developed while applying multiple sentiment analyses on data from diverse sources. However, the algorithms themselves, rather than the actual fake review identification, are the main emphasis of this research.

Our work is mainly directed to SA at the document level, more specifically, on movie reviews dataset. Machine learning techniques and SA methods are expected to have a major positive effect, especially for the detection processes of fake reviews in restaurant reviews, e-commerce, social commerce environments, and other domains. In machine learning-based techniques, algorithms such as SVM, NB, and NLP are applied for the classification purposes. SVM is a type of learning algorithm that represents supervised machine learning approaches and it is an excellent successful prediction approach. The SVM is also a robust classification approach.

The system offers a GUI environment that makes it simple to insert data into the database via forms. Web-based applications must be used. compatible with any browser (mobile browser, pc browser). A database will serve as the information store. This program's user interface is simply the standard Windows user interface; nothing more is needed. 99.9% of all new system users should be able to utilise the proposed system application without any help thanks to the proposed system's intuitive user interface. The PC on which this software will be installed must have Windows 7 and Python IDE version 1.8 or higher. Python will be installed on that Windows platform in version 3 or higher, and that is the platform used to run the specific piece of software. The Python IDE and the Microsoft SQL Server will exchange data. The frequency and fidelity of information flow inside your organisation is defined by your communication architecture. It aids in structuring your communication patterns, both within and across departments. Each business will have its own unique set of strategies, but they all need for proactive planning and investment.

II.LITERATURE REVIEW

In e-commerce, user reviews play a significant role in determining an organization's profitability. Online users read user reviews before selecting a new product or service. Because it directly affects a company's reputation and bottom line, the reliability of online reviews is crucial for organizations. As a result, some businesses use spammers to generate fake reviews. These false reviews have an impact on consumers' purchasing decisions. Numerous studies on how to spot fake reviews have been conducted in recent years. However, they still need a poll that can assess and summaries the present tactics. The task of fake review detection is described in the survey work by Ott et al.[1], which summarizes the existing datasets and their this survey thesis, we almost entirely review more than ten research papers that provide diverse approaches to effective fake user identification through the use of machine learning techniques. For each and every research topic, there is a publication outlining the benefits of seeing issues early on and a paper outlining the disadvantages of the earlier paper [2].

Customers consider online reviews carefully before making a purchase of a good or service. These are the primary sources of information regarding the features of the service we intend to acquire that come from previous customer experiences. The data set of hotel reviews is utilized in this study to introduce a number of machine learning techniques, including Naive-Bayes, Support Vector Machine, and Decision Tree, for sentiment analysis of review content and the detection of fraudulent online reviews. Sentiment analysis is currently the most exciting component of text analysis. Using sentiment analysis, we can also discern between positive and negative evaluations [3].

The suggested solution uses phase-wise processing to categories user evaluations into suspect, fraudulent, favorable, and negative categories. In this study, we use a variety of data mining approaches to process hotel evaluations. Additionally, user reviews are divided into positive and negative categories so that customers may use them to decide which products to buy. Service providers can track client opinions by carefully examining may lead to lesser demand and decrease in sales. These fake/fraudulent reviews are deliberately written to trick potential customers to promote/hype them or defame their reputations. Our work is aimed at identifying whether a review is fake or truthful one [4].

The classifiers used in our project are Support vector Machine (SVM), and Naıve Bayes. The measured results of our experiments show that the SVM algorithm outperforms other algorithms, and that it reaches the highest accuracy not only in text classification, but also in detecting fake reviews [5].

Decision Trees algorithm. : The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label [2].

Random forest: To produce a single outcome, random forest mixes the results of various decision trees. Its widespread use is motivated by its adaptability and usability because it can solve classification and regression issues [2].

Support Vector Machine Algorithm: The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyper plane is the name given to this optimal decision boundary.SVM selects the extreme vectors and points that aid in the creation of the hyper plane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method [2].

Naïve Bayes: The family of straightforward "probabilistic classifiers" known as "naive Bayes classifiers" in statistics is based on the application of Bayes' theorem with strong (naive) independence assumptions between the features. Despite being among the simplest Bayesian network models, they can reach great levels of accuracy when used in conjunction with kernel density estimation. The number of parameters required for naive Bayes classifiers is linear in the number of variables (features/predictors) in a learning problem, making them extremely scalable. Instead than using an expensive iterative approximation, which is how many other types of classifiers are trained, maximum-likelihood training can be accomplished by evaluating a closed-form expression, which requires linear time. Simple Bayes and independent Bayes are two names for naive Bayes models that can be found in statistics literature[1].

Natural Language Processing: The study of how computers interact with human language, particularly how to design computers to process and analyze massive volumes of natural language data, is known as natural language processing (NLP), a subject of linguistics, computer science, and artificial intelligence. The ultimate goal is to create a machine that can "understand" the contents of papers, including the linguistic nuances that arise from their context. After that, the system can accurately extract the knowledge and insights from the papers as well as classify and arrange the documents themselves[3].

Anaconda: Python is a mid-level object-oriented programming language that is simple to learn, easy to use, and flexible enough to do a variety of tasks (Helmus & Collis, 2016). Following its introduction in 1991 (Van Rossum & Drake Jr, 1995), its open-source nature has significantly boosted its popularity, and it is now recognised as one of the greatest programming languages to learn (Saabith, Fareez, & Vinoth raj, 2019). Python is a programming language that is accessible to everyone due to its low system requirements, free availability, and cross-platform compatibility (Mac, Windows, and Linux).It already has a sizable community made up of both regular people and eminent researchers who have produced fascinating projects in a number of different sectors, such as data science, machine learning, artificial intelligence, game and app development, and more. It is simple to find these projects by simply adding the phrase "Python" to any search due to the open-source community's constant efforts to enhance the language's capabilities. Additionally, this community provides a wealth of resources, such as tutorial videos, source codes, answers to commonly asked issues, courses, and much more. The most prevalent problems that developers run into are addressed by these tools, which are typically free to use and cover all complexity levels .[7]

NumPy: In the middle of the 1990s, a multinational team of volunteers started working on creating a data format for efficient array computation. This structure evolved into the modern N-dimensional NumPy array. The NumPy package, which comprises the NumPy array and a variety of accompanying mathematical functions, has been widely adopted in academia, government labs, and industry. It has applications in everything from gaming to space exploration. A NumPy array is a multidimensional, regular collection of elements. An array's shape and kind of components serve as its defining characteristics. An array of form (MN) holding numbers, such as complex or floating-point integers, could, for example, be used to represent a matrix. In contrast to matrices, NumPy arrays can have any number of dimensions. Additionally, they might contain a variety of things (or even a combination of things), such dates or Booleans. Actually, a NumPy array is just a handy technique to describe one or more blocks of computer memory that makes it easy to alter the numbers represented.[8]

III.CONCLUSION

Four well-known machine learning classification techniques for identifying false product evaluations will be examined using this system. Reviews that are not screened can only receive ratings like "helpful," "cool," and "funny," which means that as soon as the reviews are filtered by product, their opinions are buried and cannot be created by others.

Because an unbalanced dataset produced subpar results in our experiment, it must be handled. We discovered during the experiment that Gaussian Naive Bayes consistently produced poor test scores while SVM took the longest to train the model.

In our opinion, we cannot say that reviews got filtered by YELP recommendation system is 100% fake, because there are still other factors that may lead machine learning into false prediction. Other techniques that are potentially reliable and can be used for filtering review is using verified buyer method as some crowd source webs have been used.

An input description focused on users is transformed into a computer system using the design process. This design is essential to preventing data entry errors and demonstrating effective computer system administration for receiving accurate information. For the entry of data, it is possible to use user-friendly interfaces to handle massive volumes of data. The input design's purpose is to facilitating data entry and being error-free The information entering screen is designed to handle all data handling. It also provides options for viewing documents. The veracity of the data will be examined if they are entered. Data entry was made possible with the help of the displays. The user is given the appropriate messages indicating they are not currently in the maize. Input Design’s goal is a simple input layout to be followed.

References

[1] Estimating the incidence of fraud in online review communities: Ott M., Cardie, C., Hancock, J. 201210 is found in the book Proceedings of the 21st International World Wide Web Conference. ACM (2012)

[2] Anonymity, social image, and volunteer recruitment: a case study of the internet market for reviews, Wang, Z. B.E. Journal of Economic Analysis, 10(1), 133 (2010)

[3] Ott, M., Choi, Y., Cardie, and J.T. Hancock: Finding misleading opinion spam by whatever means necessary. In: Human Language Technologies, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA, pp. 309319 (2011)

[4] Time series-based fake opinion detection by Heydari, Tavakoli, and Salim. 58, 8392 Expert Syst. Appl (2016) Using rating behaviours, Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., and Lauw, H.W., were able to identify spammy product reviews. 19th ACM International Conference on Information Proceedings.

[5] Xie, S., Wang, G., Lin, S., Yu, P.S.: Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, pp. 823831. ACM (2012)

[6] Ye, J., Kumar, S., Akoglu, L.: Temporal opinion spam detection by multivariate indicative signals. In: ICWSM, pp. 743746 (2016)

[7]. Damien Rolon-Merette, Matt Ross, Thadde Rolon-Merette, Kinsey Church, “INTRODUCTION TO ANACONDA AND PYTHON: INSTALLATION AND SETUP”, Volume:16/Issues:05/2020.

[8]. Stefan van der Walt, S. Chris Colbert, Gael Varoquaux, “THE NUMPY ARRAY: A STRUCTURE FOR EFFICIENT NUMERICAL COMPUTION” 8 Feb-2011.