

Enhancing Trust with AI: Product Review Analysis and Segregation System

Yadagiri Rishyendra kumar
Dept. of Computer Science & Engineering
(Artificial Intelligence and Machine Learning)
MLR Institute of Technology
Hyderabad, India
rishyendra09@gmail.com

Tammysetty Tarakesh
Dept. of Computer Science & Engineering
(Artificial Intelligence and Machine Learning)
MLR Institute of Technology
Hyderabad, India
tarakeshtammysetty1@gmail.com

Guniguntla Rohith
Dept. of Computer Science & Engineering
(Artificial Intelligence and Machine Learning)
MLR Institute of Technology
Hyderabad, India
rohithgr30@gmail.com

Thalari Anurag
Dept. of Computer Science & Engineering
(Artificial Intelligence and Machine Learning)
MLR Institute of Technology
Hyderabad, India
anuragreigns366@gmail.com

Abstract — Inside the online retail space, product reviews have a significant impact on customer trust. Finding the best products or services that meet the requirements is therefore challenging. In "Enhancing Trust with AI: Product Review Analysis and Segregation System," we examine the significance of reviews for consumers when making decisions about e-commerce websites. Using sentiment analysis to direct purchase decisions, our approach separates reviews into three categories: neutral, negative, and positive sentiments. In addition, we utilize advanced machine learning and natural language processing (NLP) methods to determine if a review was written by a human or a bot. This paper analyzes various unsupervised, supervised, and semi-supervised data mining approaches for detecting fake reviews using different feature sets. Our system uses data visualization techniques to provide the data in an understandable form, further improving user convenience and comprehension.

Keywords — Fake Reviews, Sentiment Analysis, Visualization techniques, NLP techniques.

I. INTRODUCTION

The daily lives of individuals have changed as a result of the rapid expansion of the internet, which has affected their time management strategies, information consumption habits, and work environment setup. Customers primarily use reviews as a resource when deciding which products or services to use. Examples of services provided by hotels or details of products [2]. Positive customer feedback can provide a business with significant financial benefits. Because product reviews are so important, most of them are written with the false intention of exaggerating the product's efficiency, which will mislead customers into buying it. A significant number of online products are undervalued as a result of these fraudulent reviews. Furthermore, as reviews may be written by humans or automated programs, it can be difficult to identify which ones are authentic. Users publish their evaluations in a variety of ways. Some users use a single line that states something like "good," "excellent," or "perfect," while others use multiple lines [4].

Differentiating between reviews written by bots and human reviewers is an additional challenge because the content of these reviews is often complex and difficult to understand. In order to submit an experience-based review and claim ownership of a purchased product, users must have the necessary login credentials and provide proof of their previous usage of the product.

When the user posts multiple reviews for a product while maintaining a similar structure of review, we can state that it is a fraudulent review. Analyzing a wide range of variables is necessary step in identifying fake reviews [4]. Language is also an important consideration in analyzing the product reviews. Both the post's length and the particular format used for review publication can be used as markers to identify reviews that were generated by bots. Because of the explicit coding, these reviews are static and cannot fully capture the sentiment of the text. Rather, automated reviews follow preset templates, using consistent font lengths, styles, and sizes across a range of products [4].

Many important applications are based on machine learning, which is a logical trend. Machine learning's main power lies in its ability to help machines learn from their mistakes and become better versions of themselves. Algorithms for machine learning fall into three categories: supervised, semi-supervised, and unsupervised. The main objective of the many classification algorithms created for supervised learning is to choose the best classifier for our purposes. For example, a discriminant classifier called a support vector machine (SVM) divides data points using a hyperplane. Furthermore, a crucial element of the Naive Bayes classifier algorithm is the Bayes theorem [3][16].

II. LITERATURE REVIEW

Nidhi A Patel [1] has outlined the steps involved in the machine learning approach for fake review detection, including data collection, data pre-processing, feature extraction and selection, and classifier model construction and testing.

Rakibul Hassan, Md. Rabiul Islam [2] study compares the performance of three classifiers support vector machines, logistic regression, and naive Bayes with earlier semi-supervised and supervised methods in order to determine how effective each is. The suggested approach achieved remarkably high accuracy. This technique has the potential to help companies identify fraudulent reviews, allowing them to make better decisions based on real customer input.

S. Uma Maheswari, Dr.S.S.Dhenakaran [6] employed four different techniques, User-defined Fuzzy Logic, deep learning, and Machine Learning for Sentiment classification and prediction based on the accuracy and f1 score. Here the user-defined classification and fuzzy logic techniques achieved higher accuracy which is concluded by the author.

Tanjim Ul Haque and Nudrat Nawal Saber [7] used a Pool-based active learning strategy on the raw data to make the input data more accurate which parallelly improved the performance of the model with accurate results.

Using web scraping techniques, the authors, Mr. Karthikeyan T, Mr. Karthik Sekaran, and Mr. Ranjith D [8], retrieve the data from the target site and store it in an organized format. Later on, sentiment values are added to the text using preprocessing techniques like tokenizing, stemming, and bag of words on the stored data. To reduce the dimensionality of the dataset and eliminate the least significant features, LR-RFE (Logistic Regression – Recursive Feature Elimination) is utilised. In the future, machine learning algorithms and backpropagation neural networks will be used to classify data into distinct groups, such as positive, negative, and neutral.

This paper contributes to the field of detecting spam reviews by experimenting with various machine-learning algorithms.[9] Mr. Navjyotsinh Jadeja and Chirag Visani mainly focused on various methods for detecting artificially generated texts and concentrated on utilizing Twitter data. The Pattern of the reviews is analyzed by using word bigram features and syntactic components for accurate detection. The author compared SVM and Naive Bayes classifiers for the categorization of reviews as spam and not spam. Here the Naive Bayes is proven to be more accurate than SVM.

Pansy Nandwani and Rupali Verma [10] performed different approaches like Machine Learning, Lexicon-Based approaches, and Deep Learning-based approaches to detect the emotion from the text. After preprocessing the data, techniques like BOW, TF-IDF, and n-gram are used to identify the word count and extract the features. Various machine learning models, such as SVM, Decision Tree, Naïve Bayes, and deep learning models, are employed to assess the text's sentiment score and classify the data appropriately. Methods like the Transfer learning approach and hybrid approach were used, which help in reusing the pre-trained data, which indeed increases the accuracy of the mode.

The use of Part-of-Speech (POS) tagging stands out as a critical technique, according to the study "Sentiment Analysis on Online Product Review" [11] by Prof. P.S. Gaikwad, Kaushal Parmar, and Rohit Yadav. To improve the categorization of sentences as positive or negative, they utilize POS tagging to extract relevant features. This method helps greatly in distinguishing between positive and negative sentiments in comments, which makes it possible to analyse the quality of online reviews in great detail. One noteworthy application of K-means clustering in the referenced paper is its function in grouping retrieved datasets into discrete clusters. More specifically, the authors help the clustering process by using a predetermined number of clusters—in this case, two (labeled as 0 and 1). This entails grouping labelled words into appropriate clusters, which eventually separates positive and negative words.

The paper by Wenyuan Zhao [12], Chhaya Chauhan, Smriti Sehgal [27] and Purohit, Amit, Patheja, Pushpinder Singh [28] aims to make a tool that can find particular e-commerce reviews in Chinese text. They start by cleaning up the reviews, removing things like punctuation and unimportant words. Then, they divide the process into two parts: getting the text ready and teaching the system to recognize and sort the reviews. This helps the tool work better by preparing the text properly before teaching it how to identify different types of reviews. They introduced a classifier for e-commerce reviews using NBC (Naive Bayesian Classifier) and SVM (Support Vector Machine).

Astuti and Pratika [13] performed a research that involved utilizing neural network algorithms for sentiment analysis which were targeting product reviews of Amazon.com. By implementing neural network algorithms have shown accuracy of 88.2% on a dataset consisting of 1182 inputs.

Beresneva Daria [14] and Rizul Sharma [29] worked on a method of acquiring product data from various e-commerce platforms and comparing them with each other. They have tried various tools such as Scrapy, BeautifulSoup, and Selenium for data extraction. The BeautifulSoup is specifically within their system. After collecting the data, was stored in a MySQL database presented in a coherent format on a web application.

This study done by Wenyuan Zhao [15] constitutes a comprehensive systematic review of artificial content detection methods, encompassing approximately a hundred articles, of which roughly one-sixth aligned with the defined selection criteria.

In light of consumers' increasing reliance on reviews rather than just ratings, Elshirf Elmurngi [16] emphasizes the importance of textual reviews within reputation models and argues that these models should take the opinions expressed in reviews into account. This paper examines feature selection techniques for finding relevant attributes for sentiment analysis classification tasks. It highlights how important it is to select the most pertinent features to increase classification accuracy.

Eka Dyar Wahyuni and Arif Djunaidy [17] go into great detail about the evaluation procedure, list the criteria that human evaluators follow, and look into the application of the ICF and ICF++ models to detect fraudulent reviews. In addition to analyzing the level of agreement among judges during the labeling assessment process, the report looks at how to identify potential review spammers.

Mayuri Patil, Snehal Nikumbh [18] employed TF*IDF (Term Frequency-Inverse Document Frequency) algorithm for the identification of fraudulent reviews and tokenizing sentences and generating frequency matrices to compute term frequency and IDF (Inverse Document Frequency). Calculating TF-IDF values through the multiplication of values from the frequency and IDF matrices, followed by the implementation of a system controller to present authentic reviews to users subsequent to spam detection and analysis.

N Deshai, B Bhaskara Rao [19] and Chengai Sun, Qiaolin Du [26] emphasizes how important it is for e-commerce to have trustworthy review and rating systems and provides tips on how consumers can recognize and avoid misleading reviews. The main purpose of the CNN-LSTM model is to detect phony online reviews by using an embedding layer to transform words into real-valued vectors. Conversely, the primary goal of the LSTM-RNN model, which uses RNN for data extraction and LSTM for data prediction, is to detect false positives. The sequential model in the Keras deep learning Python toolkit is used to train the LSTM-RNN model.

Evaluating additional behavioral characteristics [4][23], such as those based on how often reviewers post positive or bad evaluations, how long it takes them to finish reviews, and how often they do so on a regular basis. Yin Shuqin, Feng Jing [4] proposed a method where the reviews are processed using natural language processing techniques to extract key features such as word frequency, sentiment analysis, and linguistic patterns.

Dr. V. Srividhya [30] and Aljoharah Almjawel, Sahar Bayoumi [2] has used various visualization methods such as bar charts to visualize the scraped data and compare the prices and ratings of different products across the two websites for easy understanding.

Sakshi Koli [21] conducted a study in which sentiment analysis was carried out using a variety of machine learning techniques, with an emphasis on assessing the effectiveness of various prediction models. The outcomes showed that Naive Bayes had the best precision and Support Vector Machine had the best recall.

TABLE 1. Parametric evolution of different fake review detection techniques

Author	Year	Approach	Description
Beresne va Daria	2011	Linguistic features method	A detailed examination of methods that can tell if a text is made by a person or created by a machine. The frequency counting method and the method of linguistic

			features offer effective approaches for identifying such content.
Eka Dyar Wahyuni, Arif Djunaidy	2016	Iterative Computation Framework (ICF)	The criteria used by human evaluators and the application of the ICF and ICF++ models for identifying fraudulent reviews are all covered in this paper.
Raheesa Safrin, K.R. Sharmila T.S. Shri Subangi, E.A. Vimal	2017	Feature Labelling, K-means clusters	The author examines advanced sentiment analysis methods, evaluating review sentiments with part-of-speech tagging.
Mr. Navjyotsinh Jadeja, Chirag visani,	2017	Naïve Bayes, SVM	In this paper a comparative analysis of both the algorithms is performed based on the classification of the reviews. The author finds that Naïve Bayes is more efficient than SVM.
Tanjim Ul Haque, Nudrat Nawal Saber	2018	Pool Based Active learning	The author has used a Pool-based active learning strategy on the raw data to make the input data more accurate to improve the performance of the model.
Mr. Karthikeyan T, Mr. Karthik Sekaran,	2018	Web Scraping Techniques with LR-RFE (Recursive Feature Elimination)	Website data scraped, stored, preprocessed, feature selection with LR-RFE, categorized using Back Propagation into positive, negative, neutral reviews.
Nidhi A. Patel	2018	Machine Learning Techniques	The objective of this paper is to examine several methods and strategies for identifying fraudulent evaluations. It emphasizes the various features and classifiers used for fake review detection, as well as machine learning-based techniques.
Tri Astuti, Irnawati Pratika	2019	ANN, The conjugate Scale Gradient Method	This author developed an ANN model and Conjugate method to perform sentiment analysis on product reviews with high accuracy.
Andre Sihombing A.C.M Fong	2019	Gaussian Naïve Bayes and XGBoost	Study explores Logistic Regression, SVM, Naive Bayes, and XGBoost for fake review detection, assessing effectiveness and applications.
Aljoharah Almjawel, Sahar Bayoumi,	2019	Interactive Packed bubbles, Linear chart, Stacked bars,	This document discusses the use of visualization techniques in analyzing and summarizing

Dalal Alshehr Soroor		and Word-cloud	reviews.
Yin Shuqin, Feng Jing	2019	MPINPUL (Mixing Population and Individual Nature PU Learning)	Proposed PU learning model identifies fake reviews using text, behavior, and relationship features for accurate recognition.
Wenyua n Zhao	2020	Naive Bayes Classifier, SVM	This author evaluates how well different classification models classify reviews into three groups: Positive, Negative, and Neutral.
Rakibul Hassan, Md. Rabiul Islam	2020	Supervised Machine Learning techniques	This Document introduces supervised machine learning to identify fake online reviews using TF-IDF, Empath, and sentiment features
Prof. P.S.Gaikwad, Kaushal Parmar, Rohit Yadav, Datta Supekar	2021	Web Scraping	This approach takes the product name as input, scraps the details from different sources, and displays the information on the user's window.
Mayuri patil, snehal Nikumb h	2021	Spam Detection, NLP	This paper shows Calculated TF-IDF values, multiplied frequency and IDF matrices, implemented system controller for genuine reviews post spam detection.
Uma Maheshwari, Dr.S.S. Dhenakaran	2021	Fuzzy logic, Deep Learning, and user-defined classification	Classified reviews with Fuzzy logic, user-defined classification, compared accuracy with regular DL and ML algorithms
Elshirf elmurngi	2021	Sentiment Analysis	Paper stresses textual reviews' importance in reputation models, advocating their consideration for a comprehensive understanding, surpassing mere ratings.
Pansy Nandwani, Rupali Verma	2021	Review Based Approach	Author used lexicon-based, ML, DL, and Transfer Learning for sentiment analysis and emotion detection in text, employing multiple stages.
Ahmed M.Elmo gy, Usman Tariq	2021	K-Nearest Neighbour	This paper says that the content of the reviews is processed using natural language processing to extract key features like Sentiment analysis and linguistic patterns.
N Deshai, B Bhaskar a Rao	2023	CNN- LSTM	Author used lexicon-based, ML, DL, and Transfer Learning for sentiment analysis and

			emotion detection in text, employing multiple stages.
--	--	--	---

III. PROBLEM FORMULATION

Issues with legitimacy and information overload arise when product reviews are used as an important deciding factor in online purchases. There is a need for a systematic approach that collects, classifies, and evaluates the sentiment and credibility of reviews because there is an excessive number of reviews on various platforms.

Absence of a user friendly interface to the process the huge volumes of data makes an extra obstacle that stops users from finding relevant reviews. we created a model that combines sentiment analysis, online scraping, and authenticity assessment methods, our study seeks to overcome the issues. Helping customers to make better decisions, our automated proposed model automates the process of collecting reviews on products from websites like E-commerce which helps classifying them according to their sentiment and evaluate their rightness by addressing issues. Our strategy which aims to provide customers with a convenient experience which helps in navigating the complex online product reviews.

IV. PROPOSED SOLUTION

We divided our approach into three steps for easier understanding. It starts with data collection and web scraping, then moves towards the sentiment analysis where we are using machine learning algorithms on reviews for classification. And then finally ends with presenting the results in an understandable way for decision-making which is led by the customer..

STEP 1: DATA COLLECTION AND WEB SCRAPING

- Copying the product link and paste it on our interface
- Performing web scraping using the BeautifulSoup library
- Gathering reviews for different e-commerce websites
- Storing the gathered data in a secured format

STEP 2: SENTIMENT ANALYSIS AND REVIEW CATEGORIZATION

- Comparing reviews using resources such as NLTK and Scikit-learn
- Classifying the reviews into three categories, Positive, Negative, and Neutral by using various Machine Learning tasks like clustering, and classification.
- Identifying the patterns from the classified data and proceeding to the next step.

Scikit-learn, on the other hand, is a toolkit that offers a variety of machine-learning tools and algorithms for tasks like classification, regression, clustering, and more. Using labeled data, machine learning algorithms are created in this

step to detect patterns and produce a state of recently published reviews. Afterward, the reviews are divided into three groups based on the implicit opinions expressed in them.

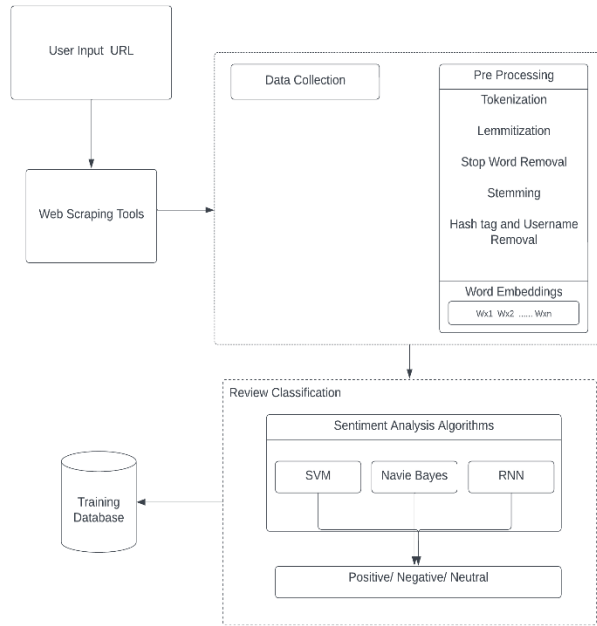


Figure 1: Detailed architecture for classifying reviews based upon sentiment analysis for the training data

STEP 3: DATA PRESENTATION AND USER INTERFACE:

In the final step, our system focuses on presenting the data that is analyzed, in a user-friendly interface that can be used for easy decision-making.

A well-designed user interface has been developed that involves users effectively and provides insights from the sentiment analysis and review categorization process. Data presentation involves the creation of interactive dashboards and visualizations by using visualization libraries such as Matplotlib, Seaborn, or Plotly from which the user can archive the overview of the product reviews.

- Creating interactive dashboards and visuals using Visualization libraries such as Matplotlib, seaborn, and plotly.
- Presenting the data in the user-understandable format.
- Providing customizable filters based on opinion of the user.

The system initiates sentiment analysis in response to this interaction. Users can then visually understand the distribution of sentiments across reviews thanks to the system's insightful visual representations that it generates in response. This well-rounded strategy makes it simple for users to interact with the product reviews and provides insightful information, which helps people who are looking for genuine and insightful feedback make better decisions.

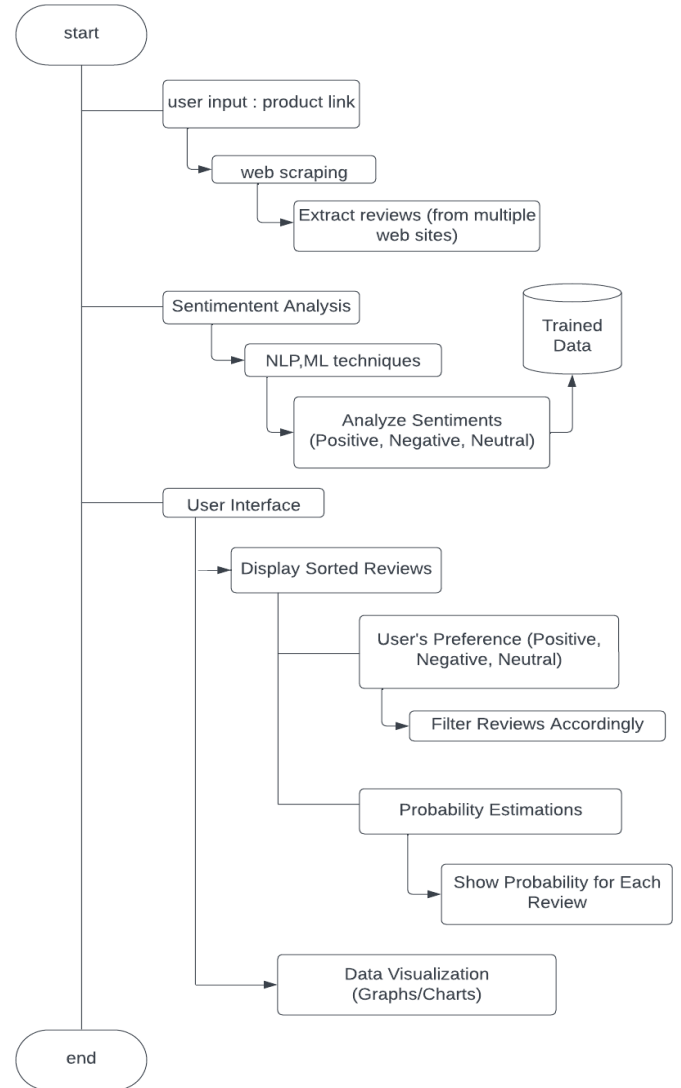


Figure 2: Analyzing the test data reviews and visualizing them after classifying

V. CONCLUSION

Our research has demonstrated how technology can revolutionize customer decision-making in the e-commerce space. Our model accelerated review aggregation, sentiment analysis, and authenticity assessment by combining web crawling, machine learning, and natural language processing. This approach provides visitors the ability to confidently and quickly browse through a variety of reviews. The advantages are obvious: better understanding, more trust, and well-informed decisions. Future developments in machine learning for more complex sentiment analysis and UI improvements should improve user experiences. This study highlights how technology is changing the way that e-commerce reviews are analyzed, providing insight into a future in which customers have even more power and confidence when making purchases online.

REFERENCES

- [1] Nidhi, A. Patel, "A Survey on Fake Reviews Detection using Machine Learning Techniques", *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, doi:10.1109/CCAA.2018.8777594
- [2] Rakibul Hassan, Md. Rabiul Islam, "A Supervised Machine Learning Approach to Detect Fake Online Reviews", *2020 23rd International*

Conference on Computer and Information Technology (ICCIT), 19-21 December, 2020, doi:10.1109/ICCIT51783.2020.9392727

- [3] Andre Sihombing, A.C.M. Fong, "Fake Review Detection on Yelp Dataset Using Classification Techniques in Machine Learning", *2019 International Conference on contemporary Computing and Informatics (IC3I)*, doi:10.1109/IC3I46837.2019.9055644
- [4] Yin Shuqin, Feng Jing, "Fake Reviews Detection Based on Text Feature and Behaviour Feature", *2019 IEEE 21st International Conference on High Performance Computing and Communications*, doi:10.1109/HPCC/SmartCity/DSS.2019.00277
- [5] Aljoharah Almjawel, Sahar Bayoumi, Dalal Alshersh, Soroor Alzahrani, Munirah Alotaibi, "Sentiment Analysis and Visualization of Amazon Book's Reviews", *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, DOI: 10.1109/CAIS.2019.8769589
- [6] S. Uma Maheswari, Dr. S.S. Dhenakaran, "Detection of fake and Genuine Reviews with Hybridization of Fuzzy and Neural Networks Techniques", June 2021
- [7] Tanjim Ul Haque, "Sentiment Analysis on Large Scale Amazon Product Reviews", 2018
- [8] Mr. Karthikeyan T, Mr. Karthik Sekaran, Mr. Ranjith D, Mr. Vinoth Kumar V, Mr. Balajee, "personalized content extraction and text classification using effective web scrapping techniques", 2019
- [9] Mr. Navjyotsinh, chirag visani, "A study on different machine learning techniques for spam reviews detection", 2017
- [10] Pansy Nandwani, Rupali Verma, "A review on sentiment analysis and emotion detection from the text", 2021.
- [11] Prof. P.S. Gaikwad, Kaushal Parmar, Rohit Yadav, Datta Supekar, "Implementation of web scrapping for e-commerce website", 2021.
- [12] Raheesa Safrin, K.R.Sharmila, T.S.Shri Subangi, E.A.Vimal, "Sentiment analysis on online product review", 2017.
- [13] Tri Astuti, Inawati Pratika, "Product Review Sentiment Analysis by Artificial Neural Network Algorithm", 2019.
- [14] Beresneva Daria, "Computer-generated Text Detection Using Machine Learning: A Systematic Review", 2011.
- [15] Wenyuan Zhao, "Classification of Customer Reviews on Ecommerce Platforms Based on Naive Bayesian Algorithm and Support Vector Machine", 2020.
- [16] Elshirf Elmurngi, "Fake Reviews Detection using Supervised Machine Learning (IJACSA) International Journal of Advanced Computer Science and Applications", Vol. 12, No. 1, 2021.
- [17] Eka Dyar Wahyuni, Arif Djunaidy, "fake review detection from a product review using modified method of iterative computation framework".
- [18] Mayuri Patil, Snehal Nikumbh, "Fake Product Review Monitoring and Removal for Genuine Product Reviews", (International Journal of Scientific Research and Engineering Development— Volume 4 Issue 3, May -June 2021.
- [19] N Deshai, B Bhaskara Rao, "Deep Learning Hybrid Approaches to Detect Fake Reviews and Ratings", (Journal of Scientific & Industrial Research Vol. 82, January 2023).
- [20] Ahmed M.Elmoogy, Usman Tariq, "Fake Reviews Detection using Supervised Machine Learning", ((IJACSA) International Journal of Advanced Computer Science and Applications Vol. 12, No. 1, 2021)
- [21] Koli. S, "Sentiment Analysis with Machine Learning Techniques and Improved J48 Decision Tree Technique", International Journal of Computer Sciences and Engineering, Vol. 9, Issue.6, June 2021.
- [22] Vinodhini, G., & Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 6, June 2012.
- [23] Zheng. B., Thompson. K., Lam. S., S. Yoon "Customers Behavior Prediction Using Artificial Neural Network."
- [24] Rutuja B.Ardak, Prof. Girish S. Thakare, "A review on fake product review detection and removal techniques", Volume 9, Issue 8 August 2021
- [25] Anusha Sinha, Nishant Arora, "- Fake Product Review Monitoring Using Opinion Mining (International Journal of Pure and Applied Mathematics", Volume 119 No. 12 2018.
- [26] Chengai Sun, Qiaolin Du, "Exploiting Product Related Review Features for Fake Review Detection", 2016.
- [27] Chhaya Chauhan, Smriti Sehgal, "Sentiment analysis on product reviews", *2017 International Conference on Computing, Communication and Automation (ICCCA)*, doi: 10.1109/CCAA.2017.8229825.
- [28] Purohit, Amit, Patheja, Pushpinder Singh, "Product review opinion based on sentiment analysis", 2023, doi: 10.3233/JIFS-213296
- [29] Rizul Sharma, "Web Data Scrapping", March 2020, DOI: 10.13140/RG.2.2.15125.76009
- [30] Dr. V. Srividhya, "Scraping and Visualization of Product Data from E-commerce Websites", May 2019.

