INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 1882-1886

e-ISSN : 2583-1062

Impact Factor: 5.725

ENHANCING ACCURACY IN PREDICTING FORGERY ANALYSIS COMPARISON OF E- COMMERCE APPLICATIONS

Marcus Antony R[1], Dr. Vaidehi V[2]

[1]PG Student, Department of Computer Applications, Dr. M.G.R Educational and Research Institute
Chennai, Tamil Nadu, India.

[2]Professor, Department of Computer Applications, Dr. M.G.R Educational and Research Institute
Chennai, Tamil Nadu, India.

ABSTRACT

In the era of seamless online shopping, where consumer decisions are heavily influenced by online product reviews, the proliferation of fake or misleading feedback poses a significant challenge. This research paper proposes an innovative approach to tackle this issue by employing Machine Learning (ML) algorithms to detect and filter out fake reviews from online review feeds.

Leveraging the wealth of data available in online review platforms, the proposed methodology aims to enhance consumer trust and confidence in online purchasing decisions. The study explores various ML techniques, including natural language processing (NLP) and sentiment analysis, to analyze review text and metadata for patterns indicative of fraudulent behaviour.

Through extensive experimentation and evaluation on real-world datasets, the effectiveness and accuracy of the proposed approach are demonstrated. The findings highlight the potential of ML algorithms in mitigating the detrimental effects of fake reviews on consumer trust and brand credibility in the digital marketplace. Additionally, implications for businesses and online platforms are discussed, emphasizing the importance of implementing robust review validation mechanisms to ensure the integrity of consumer feedback.

Keywords: Online product reviews, Fake reviews, Machine Learning algorithms, Natural Language Processing, Sentiment Analysis, Consumer trust, Digital marketplace

1. INTRODUCTION

In recent years, there has been significant attention towards detecting fake consumer reviews, attributed to the surge in online purchases. Conventional methods for identifying fake reviews typically rely on factors such as review content, product details, and reviewer characteristics. To enhance the accuracy of fake review detection, two neural network models are proposed, integrating traditional bag-of-words representations with word context and consumer emotions. These models are engineered to learn document-level representations through three distinct sets of features: n-grams, word embeddings, and various lexicon-based emotion indicators [1].

As e-commerce continues to advance rapidly, providing a platform for buying and selling products and services online, customers are increasingly turning to these digital marketplaces to fulfil their needs. Post-purchase, customers often express their experiences, sentiments, and emotions regarding the products and services they've acquired through reviews. These reviews serve as valuable feedback for other potential buyers, offering insights into the quality, usability, and overall satisfaction associated with the purchased items. As such, understanding and analyzing customer reviews has become an essential aspect of e-commerce management, allowing businesses to gauge customer satisfaction levels, identify areas for improvement, and tailor their offerings to better meet customer needs and preferences [2].

In e-commerce, user reviews hold considerable sway over organizational revenue, shaping consumer decision-making processes. The credibility of these reviews is pivotal for maintaining business reputation and financial success. However, the proliferation of fake reviews, often incentivized by businesses paying for fraudulent endorsements, erodes consumer trust. To address this, a survey paper delves into fake review detection methods and datasets, providing a comprehensive analysis. It critically evaluates traditional statistical machine learning and deep learning approaches. Additionally, benchmarking experiments showcase RoBERTa's superior performance, suggesting its potential as a future benchmark. [3].

Due to the diversity and intricacy of models and features associated with Automatic Fake News Detection, our systematic review diverges from existing literature by concentrating solely on content-based models and features, excluding hybrid approaches. Following a structure akin to other sector reviews, our aim is to facilitate comparison among them. The anticipated outcome is to offer guidance to researchers tackling content-based fake news detection and to aid system developers in implementing precise content-based fake news detectors [4].

As customer reliance on reviews grows for purchasing decisions, particularly on e-commerce and social networking platforms, the authenticity of each review becomes paramount. Past research has introduced a range of machine learning techniques to discern false product reviews. Identifying the most suitable machine learning algorithm for a specific dataset type is crucial. The results demonstrate that the systems exhibit robust classification performance across all datasets, irrespective of sentiment polarity or product category [5].

2. LITERATURE SURVEY

Hajek et al. (2020) conducted a study wherein they employed two distinct neural network models, namely the deep feed-forward neural network (DFFN) and convolutional neural network (CNN), utilizing the Amazon product review dataset as the basis for their analysis. T

he researchers focused on extracting feature sets, which included word emotions and N-grams, from the dataset to enhance the accuracy of their models. Their methodology involved training these neural network models on the extracted features to classify reviews as genuine or fake. The results of their study revealed promising accuracy rates, with the DFFN method achieving an accuracy of 82% and the CNN method, thereby contributing to the advancement of research in the field of online review authenticity assessment [6].

Alsubari et at. (2020) presented a study of Linguistic Inquiry and Word Count (LIWC) emerged as a prominent analysis tool utilized by researchers. This study demonstrated improved results in fake review identification by incorporating parts of speech (POS) features into the LIWC framework. Examples of LIWC output features include self-reference (e.g., "I," "my," "me"), positive emotions (e.g., "love," "nice," "sweet"), negative emotions (e.g., "hurt," "ugly," "nasty"), social words (e.g., "talk," "mate," "they," "child"), big words (words with more than six letters), overall cognitive words (e.g., "cause," "know," "ought"), articles (e.g., "a," "an," "the"), and LIWC summary variables scores such as authenticity, clout, cognitive words, and social words enables researchers to gain deeper insights into the linguistic characteristics of textual data, facilitating more accurate and nuanced analysis in various research domains [7].

Zhang et al. (2023) proposed a novel end-to-end framework that has integrating behavioral and textual information for fake reviewer detection. This framework includes a behavior-sensitive feature extractor and a context-aware attention mechanism. Rigorous evaluation on real-world datasets from Yelp.com demonstrates that this method achieves state-of-the-art results in fake reviewer detection. This advancement signifies a step towards reducing the human labor costs associated with detecting fake reviewers in e-commerce platforms [8].

Shunxiang et al. (2023) proposes the SIPUL model for detecting fake reviews, capable of continuously learning from streaming data. Initially, reviews are divided into subsets based on sentiment intensity, distinguishing between strong and weak sentiment sets. Positive and negative samples are then extracted using SCAR and Spy technology. A semi-supervised PU learning detector is constructed from these initial samples to iteratively identify fake reviews in the data stream. Updates to both the initial samples and the PU learning detector are made based on detection results. Historical record points guide the ongoing deletion of old data to maintain a manageable training sample size and prevent overfitting [9].

Shukla et al. (2024) proposed a pioneering solution to tackle the pervasive problem of fake reviews in online commerce. In response to the significant damage caused by fraudulent reviews to businesses and consumer welfare, Anup Singh introduced a novel approach centered around digital identity verification.

This innovative strategy aims to verify a user's identity through various forms of digital information not previously associated with online reviews. By highlighting the limitations of existing techniques, Anup Singh underscores the potential of digital identity verification as a promising solution to combat fake reviews effectively. Through this case study, Anup Singh aims to offer insights into the benefits and challenges associated with this approach and its effectiveness in addressing the fake review problem plaguing online commerce platforms [10].

3. METHODOLOGY

In this study, a systematic approach was employed. Initially, a dataset containing both fake and truthful reviews was compiled. Subsequently, various classification algorithms were applied to the dataset. Figure 1 illustrates the methodology utilized for conducting this research.

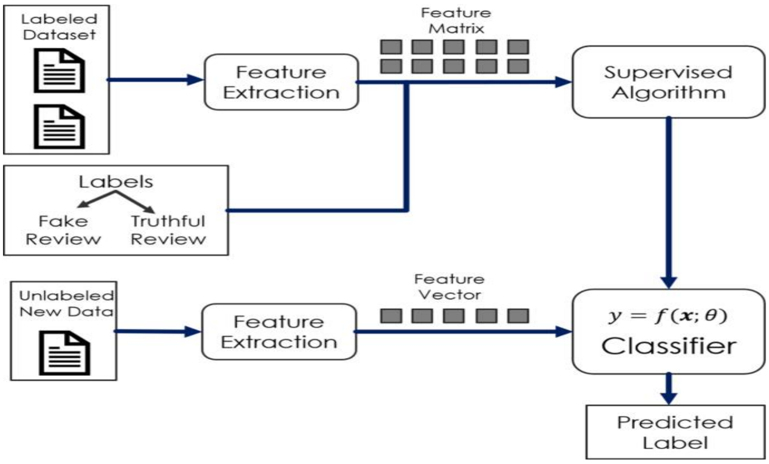


Figure 1. Methodology of the Proposed System

A. Dataset:

For this study, a Kaggle dataset was utilized, accessible through a sub-link on Kaggle's website (https://www.kaggle.com/general/243411). This dataset provided a comprehensive range of attributes to analyze, including Verified Purchase, Product Title, Review Title, Rating, Review Text, and classification into fake or real reviews. Exploratory analysis of the dataset was conducted to identify trends within the database. It comprises 10 features and consists of 21,000 instances, evenly distributed between 10,500 fake and 10,500 real reviews. To facilitate analysis, the data was transformed into numerical format, as many machine learning algorithms require numerical inputs.

B. Algorithms Used:

- XG Boost : Extreme Gradient Boosting, is a machine learning algorithm that implements gradient boosted decision trees, prioritizing speed and performance. It incorporates L2 regularization to enhance model generalization. In our study, we applied XGBoost in two distinct manners. Initially, we utilized it on non-text features such as sentiment polarity, rating, and review length, achieving an accuracy of 57%. Subsequently, we employed XGBoost on text data represented as the CSR IDF matrix, resulting in improved performance with an accuracy of 64%.
- Random Forest : Random Forest is an ensemble learning method comprising multiple decision trees that collectively make predictions based on class votes. In our study, we enhanced the model by combining non-text features with the CSR matrix, using them as additional column features. This combined dataset was then utilized with a Random Forest Classifier consisting of 1000 trees and a maximum depth of 4. Bootstrapping was employed instead of using the entire dataset, and the Gini index guided the tree splitting process. The resulting model achieved an accuracy of 65.93% on the combined features dataset.
- SVM : Support Vector Machine (SVM) is a supervised learning model developed by Catanzaro, utilized for both regression analysis and categorization tasks. SVM is known for its resilience and is constructed within a statistical learning framework. In our study, we applied SVM linear classifier to two types of datasets: review content and non-text features. For the text dataset, we achieved an accuracy of 62.33%, while for other features such as review length and sentiment polarity, the accuracy was 54.1%. We employed the RBF kernel while training on non-text features.
- Logistic Regression : Logistic regression is employed when the dependent variable is binary, yielding predictive outcomes. This model is utilized to investigate the relationship between independent variables of nominal, interval, ordinal, or ratio levels and a binary dependent variable.

4. RESULTS AND DISCUSSION

An exhaustive examination of the dataset was conducted, and diverse results were obtained during the testing phase to assess the accuracy of the model. This process aimed to formulate a system capable of detecting the fake product review based on the gathered insights.

Evaluation Metrics

Table 1: Algorithm vs. Accuracy Result in Percentage

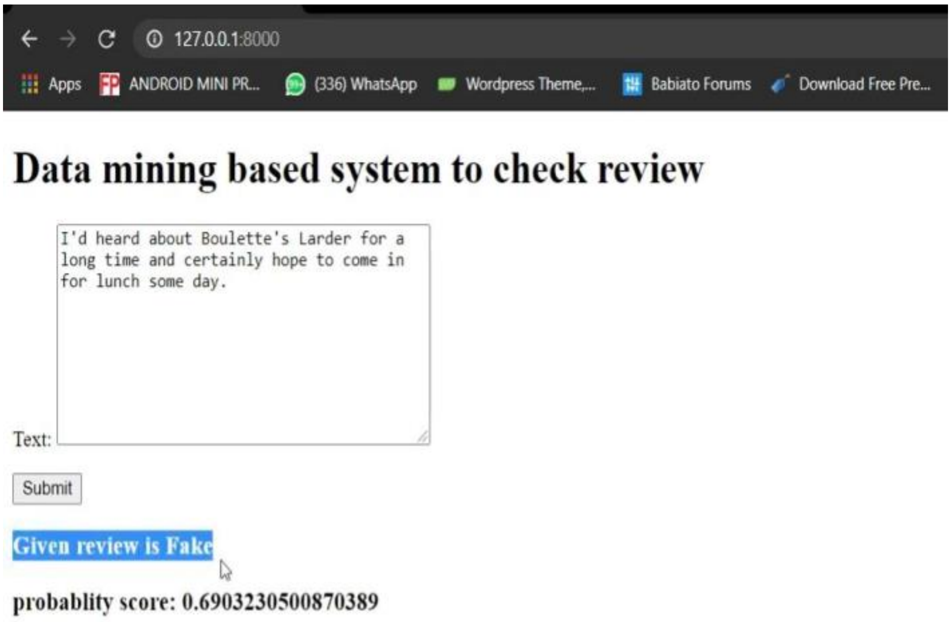
ALGORITHM	ACCURACY (%)
XG Boost	64.00

Support Vector Machine (SVM)	72.00
Logistic Regression (LR)	56.27
Random Forest (RF)	65.93

The above table shows the result of Accuracy between different algorithms in percentage. It is clear that SVM algorithm has the greatest accuracy of all.

SCREENSHOTS

The result includes a website which gives asks user to enter a text and then it predicts the review as a real or a fake .



Data mining based system to check review

Text: I'd heard about Boulette's Larder for a long time and certainly hope to come in for lunch some day.

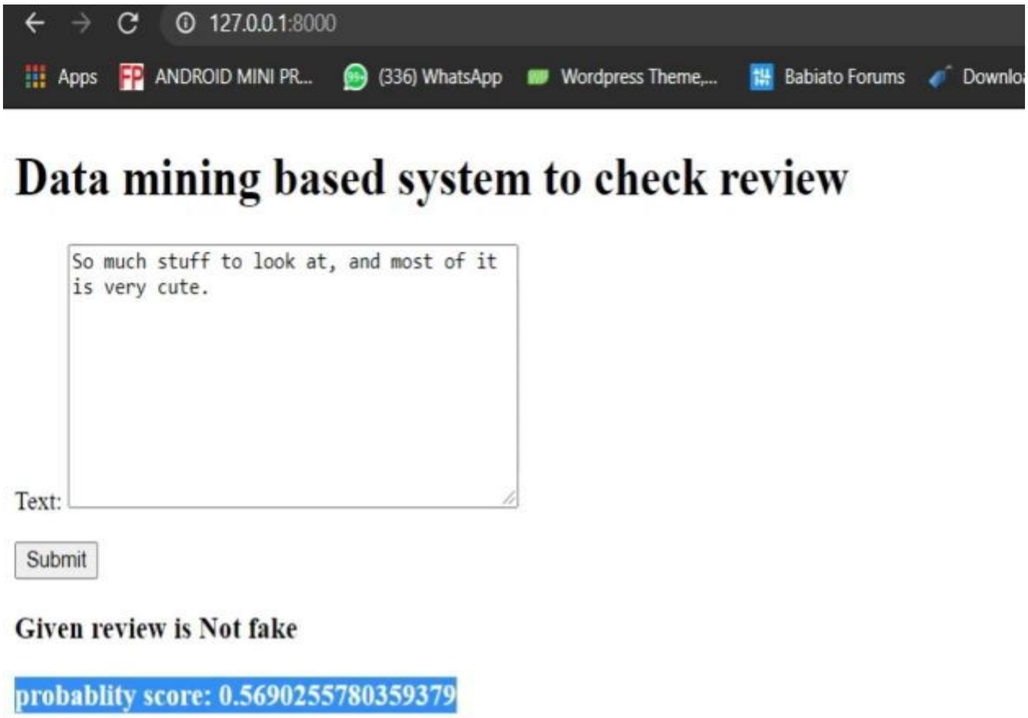
Submit

Given review is Fake

probablity score: 0.6903230500870389

Figure 2. Given Review is Fake

In Figure 2, Some fake review texts has been entered into the Text box. After clicking on the Submit button, it process the given text and shows the result as “Given review is Fake”.



Data mining based system to check review

Text: So much stuff to look at, and most of it is very cute.

Submit

Given review is Not fake

probablity score: 0.5690255780359379

Figure 3. Given Review is not Fake

In Figure 3, The entered text inside the text box is processed and returns the result as “Given Review is Not Fake”.

5. CONCLUSION

Contrary to our initial expectations, the inclusion of reviewer-centered features did not notably enhance the performance of our models. This outcome may be attributed to the limited availability of reviewer data, as many reviewers only had a single review record. Consequently, the majority of reviewer-centered features lacked robust definition, such as those based on the maximum number of reviews in a day or the standard deviation of ratings. Interestingly, our study revealed that simplicity often prevails in achieving optimal results. By utilizing a Support Vector Machine with carefully tuned parameters, we attained an impressive accuracy rate of nearly 72%. This outcome surpassed the performance of all other classifiers tested in the study. Thus, our findings underscore the importance of methodological simplicity and parameter optimization in achieving superior model performance, even in scenarios where more complex feature engineering is anticipated.

6. REFERENCES

[1] Hajek, P., Barushka, A. and Munk, M., (2020), " Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining", Neural Computing and Applications, Vol. 32(23), pp.17259-17274.

[2] Alsubari, S.N., Deshmukh, S.N., Al-Adhaileh, M.H., Alsaade, F.W. and Aldhyani, T.H., (2021) "Development of integrated neural network model for identification of fake reviews in E-commerce using multidomain datasets" , Applied Bionics and Biomechanics, pp.1-11.

[3] Mohawesh, R., Xu, S., Tran, S.N., Ollington, R., Springer, M., Jararweh, Y. and Maqsood, S., (2021) "Fake reviews detection: A survey" , IEEE Access, Vol 9, pp.65771-65802.

[4] Capuano, N., Fenza, G., Loia, V. and Nota, F.D., (2023) " Content-based fake news detection with machine and deep learning: A systematic review" , Neurocomputing, Vol. 530, pp.91-103.

[5] Singh, Y., Teotia, R., Kadian, K., Garhwal, S. and Dwivedi, V., (2024), "Fake Review Detection and Removal: A Comparative Analysis using ML and DL Models" , Grenze International Journal of Engineering & Technology (GIJET), Vol. 10(1).

[6] P. Hajek, A. Barushka and M. Munk (2020), "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," Neural Computing and Applications, vol. 32, no. 23, pp. 17259–17274.

[7] Alsubari S. N., Shelke M. B., and Deshmukh S. N., (2020), "Fake reviews identification based on deep computational linguistic features," International Journal of Advanced Science and Technology, vol. 29, no. 8s, pp. 3846–3856.

[8] Zhang, D., Li, W., Niu, B. and Wu, C.,(2023), "A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information", Decision Support Systems, Vol. 166, p.113911.

[9] Shunxiang Z., Aoqiang Z., Guangli Z., Zhongliang W., and KuanChing L.,(2023), "Building Fake Review Detection Model Based on Sentiment Intensity and PU Learning," in IEEE Transactions on Neural Networks and Learning Systems, Vol. 34, no. 10, pp. 6926-6939.

[10] Shukla, A.D. and Goh, J.M., (2024), " Fighting fake reviews: Authenticated anonymous reviews using identity verification" , Business Horizons, Vol. 67(1), pp.71-81.