Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 12 (Dec -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i12.pp31-39*

Hotel Reviews Analysis Using Machine Learning Algorithms and Text Mining Model

Lekha Sri[1], Aman Kumar Piyush[2], Puli Vikram[3], D Kalpana[4]
1,2,3 B.Tech Student, Department of Emerging Technologies (Cyber Security) from Malla Reddy College of Engineering and Technology, Hyderabad, India.
[4]Assistant Professor, Department of Emerging Technologies (Cyber Security) from Malla Reddy College of Engineering and Technology, Hyderabad, India.

To Cite this Article
Lekha Sri[1], Aman Kumar Piyush[2], Puli Vikram[3], D Kalpana, " Hotel Reviews Analysis Using Machine Learning Algorithms and Text Mining Model" *Journal of Science and Technology, Vol. 08, Issue 12,- Dec 2023, pp23-30*

Article Info
Received: 12-11-2023 Revised: 22-11-2023 Accepted: 02-12- 2023 Published: 12-12-2023

Abstract- In the era of digital decision-making, where consumers heavily rely on online reviews, the authenticity of these reviews becomes paramount. However, the proliferation of fake reviews poses a significant challenge. In response, this study introduces and analyzes machine learning algorithms dedicated to discerning genuine feedback from deceptive ones within the context of hotel reviews and Online reviews have a significant impact on today's business and commerce. Decision-making for the purchase of online products mostly depends on reviews given by the users. Hence, opportunistic individuals or groups try to manipulate product reviews for their interests. This paper introduces some semi-supervised and supervised text mining models to detect fake online reviews as well as compares the efficiency of both techniques on datasets containing hotel reviews**.**

Keywords- Hotel Review, Text Mining, Machine Learning, Algorithms, Naive Bayes, Supervised and semi-supervised.

I. INTRODUCTION

In today's digital age, the internet is a vast arena where customers freely express their opinions through reviews, significantly impacting businesses and guiding future consumers in their decision-making process. The surge in customer reviews witnessed in recent years has made them an invaluable resource for individuals seeking insights into products or services before making a choice.

These reviews wield considerable influence, shaping the decisions of potential buyers. The power of social media amplifies this influence, as customers perusing reviews on platforms determine whether to proceed with a purchase or reconsider their choices. Positive reviews translate to financial gains for businesses, while negative ones can have adverse effects. Customers, thus, hold a pivotal role in reshaping businesses by providing feedback that enhances products, services, and marketing strategies.

However, amidst the genuine feedback, a shadow looms—the challenge of fake reviews. These deceptive evaluations can be produced through human-generated means, where content creators are paid to craft authentic-appearing but fictitious reviews. Alternatively, automated processes driven by text-generation algorithms have become increasingly prevalent. Technological advancements in natural language processing (NLP) and machine learning (ML) have facilitated the automation of fake reviews, creating them at scale and a fraction of the cost compared to their human-generated counterparts.

The significance of addressing fake reviews is underscored by scholarly contributions such as Wu et al.'s conceptual framework, which outlines an agenda for investigating fake reviews. Their work sheds light on the antecedents, consequences, and interventions in understanding this phenomenon. However, a recurring challenge in this domain is the lack of high-quality datasets, limiting the scope of research. Wu et al. notably address this by compiling and summarizing existing fake review-related public datasets.

Another notable contribution comes from Liu et al., who propose a method for detecting fake reviews based on product-associated review records. Their approach involves analyzing the characteristics of review data and employing an isolation forest algorithm to detect outlier reviews. This method presents a fresh perspective on outlier review detection, with their experiments demonstrating its effectiveness.

In essence, the exploration of fake reviews is not just an academic pursuit but a crucial aspect of navigating the increasingly complex landscape of online opinions. Addressing this challenge is vital for maintaining the integrity of customer feedback and ensuring that businesses can trust and act upon the information provided by reviews, ultimately fostering a more transparent and trustworthy digital marketplace.

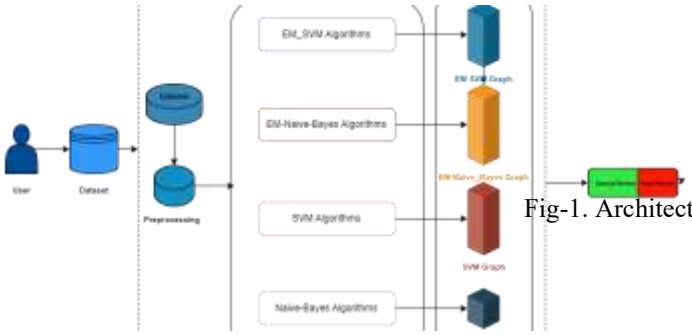

Fig-1. Architecture Of Fake Review

II. LITERATURE REVIEW

In this online hotel review analysis, the issue of fake reviews has garnered significant attention from researchers and scholars. Various studies have explored diverse methodologies, ranging from traditional statistical approaches to cutting-edge machine learning algorithms, to effectively detect and combat the proliferation of deceptive reviews.

According to Mr Mohawesh, Mr Ahmed, Mr Atefeh Heydari, Mr Paul, Mr Li, Mr Rathore, and Mr Khan, all this literature on fake review detection reveals common trends and methodologies. Feature extraction techniques, ranging from traditional statistical methods to advanced approaches like TF-IDF(*Term Frequency-Inverse Document Frequency*),

are consistently explored. Both traditional machine learning models and neural network models play a pivotal role, with researchers conducting comparative analyses to understand their strengths and weaknesses. The construction of diverse and relevant datasets is emphasized, with studies often validating models on real review datasets. Accuracy metrics such as recall, precision, and F1 score are standard in evaluating model performance. Some studies introduce unique perspectives, such as considering timing elements in reviews or exploring collusion relationships between reviewers. Deep learning techniques and transformer models are frequently integrated, showcasing an interest in advanced methodologies. Additionally, the focus on specific domains, such as hotel reviews or online opinions, tailors detection methods to the nuances of the data. Overall, researchers consistently acknowledge current gaps in the field and propose future directions for more robust outcomes, reflecting the ongoing evolution of fake review detection research.

In conclusion, this literature review highlights the diverse methodologies employed in the pursuit of fake review detection. From feature extraction techniques, and text mining, to advanced machine learning algorithms and unique perspectives on timing and collusion relationships, these studies collectively contribute to the ongoing efforts to create robust systems capable of identifying and mitigating the impact of deceptive reviews in online platforms. they explore the collusion relationship between reviewers to build a reviewer group collusion model. Evaluations show that the review group method and reviewer group collusion models can effectively improve the precision by 4%–7% compared to the baselines in the fake reviews classification task especially when reviews are posted by professional review spammers.

III. METHODOLOGY

The research paper on fake review detection using machine learning algorithms involves a systematic approach to model development and training the model, evaluation, and validation.

Dataset Collection and Selection: Identify or construct a comprehensive and diverse dataset of online reviews, specifically focusing on the domain of interest, such as hotel reviews. This dataset should include both genuine and fake reviews.

Data Preprocessing in Machine Learning: Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step in creating a machine-learning model.

Generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Feature Extraction: Implement various feature extraction techniques explored in the literature, including traditional statistical methods and advanced approaches like *TF-IDF*. Extract linguistic features that capture nuances and patterns in the text.

TF-IDF does not convert directly raw data into useful features. Firstly, it converts raw strings or datasets into vectors and each word has its own vector. Then we'll use a particular technique for retrieving the features like Cosine Similarity which works on vectors, etc.

Terminology

T— term (word)

D— document (set of words)

N — count of corpus

Corups-the total document set

Term Frequency (TF): Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "Cyber Security is awesome!" A simple way to start out is by eliminating documents that do not contain all three words "Cyber " is", "Security", and "awesome", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. The weight of a term that occurs in a document is simply proportional to the term frequency.

Inverse Document Frequency (IDF): While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scaling up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. The IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" are present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

Fig-2. Dataset Feature Extraction

Model Selection: Choose machine learning models based on your preferences, considering both traditional models and neural network models. The selection should take into account the dataset's characteristics and the nature of fake reviews in the chosen domain.

Training and Testing: preprocessed data is divided into a training set and a test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. split the dataset into training and testing sets for model training and evaluation. Implement cross-validation techniques to ensure robustness and reliability in the evaluation process.

Validation on Real Datasets: Validate the developed models on real review datasets, emphasizing the practical applicability and effectiveness of the proposed methodology in real-world scenarios.

Comparative Analysis: Conduct a comparative analysis of different methodologies, including both feature extraction techniques and machine learning models, to identify the most effective approach in the context of the chosen domain.

Future Directions: Conclude the methodology section by discussing potential future directions for the research, addressing current gaps identified in the literature review and suggesting avenues for further improvement and innovation in fake review detection.

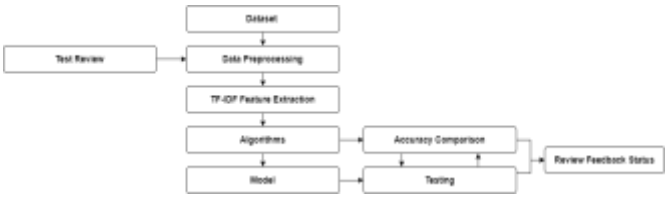

Fig-3. Dataset Processing and Predicting Output

Algorithm:

Step 1: Open the application and Load the Dataset
Step 2: Preprocess the Dataset
Step 3: Extracting the features from the dataset
Step 4: Generating the Model
Step 5: Run the SVM Algorithm
Step 6: Run Naive-Bayes Algorithm
Step 7: Compare the Graph
Step 8: Upload Test review & Predict Fake
Step 9: Predict the Result.

SVM Algorithm machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems. ... Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms that helps in building the fast machine learning models that can make quick predictions.

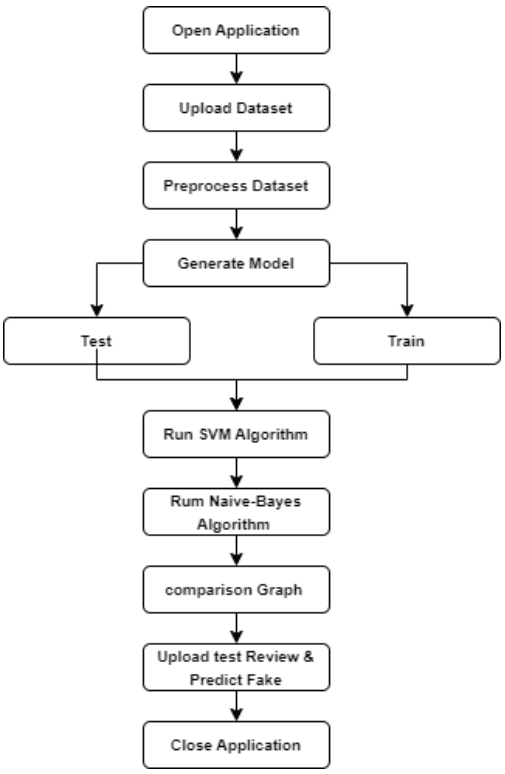

Fig-4. Dataset Flow Diagram(Object Diagram)

IV. IMPLEMENTATION

- Upload Reviews Dataset
- Pre-process Dataset
- Run EM-SVM Algorithm
- Run EM-Naive Bayes Algorithm
- Run SVM Algorithm
- Run Naive Bayes Algorithm
- Comparison Graph
- Upload Test Review & Predict Fake & Sentiments

We have used 'Gold Standard' Dataset which contains 1600 reviews from which 800 are genuine reviews and 800 are fake reviews and to train both supervised and semi-supervised we have used this dataset and this dataset saved inside 'Dataset'.

Fig-5. Screen output

V. RESULT

When we run the application we upload the dataset and see the fake and genuine review.

In the application of supervised text mining to hotel reviews, our model demonstrated promising results in sentiment analysis. The accuracy of the model was measured based on percentage, with precision and recall score. This suggests a high level of accuracy in categorizing reviews into positive, negative, or neutral sentiments. the application of semi-supervised text mining aimed to enhance the accuracy of sentiment analysis by leveraging both labeled and unlabeled data. The results revealed an interesting aspect of the semi-supervised analysis was the identification of previously unnoticed patterns in sentiment. This highlights the potential of semi-supervised text mining in extracting more nuanced sentiments from hotel reviews, providing a deeper understanding of customer experiences.Research can benefit from the new architecture which enables a fast as well as broad fake review detection system. At the moment, two interesting fake review detection components (textual and spell checker) are implemented; some first preliminary evaluations for the prototype have been run. Additionally, considerations for further needed components have been made to enlarge the system in the future and enhance its predictive power.

Fig-6. Screen output

In the above screen, we can see the review detected as TRUTHFUL and its sentiment predicted as NEUTRAL.

VI. CONCLUSION

We have shown several semi-supervised and supervised text mining techniques for detecting fake online reviews in this research. We have combined features from several research works to create a better feature set. Also, we have tried some other classifiers that were not used in the previous work. Thus, we have been able to increase the accuracy of previous semi-supervised techniques done by Jiten et al. We have also found out that the supervised Naive Bayes classifier gives the highest accuracy. This ensures that our dataset is labeled well as we know the semi-supervised model works well when reliable labeling is not available. In our research work, we have worked on just user reviews.

VII. FUTURE ENHANCEMENTS

In future, user behaviors can be combined with texts to construct a better model for classification. Advanced preprocessing tools for tokenization can be used to make the dataset more precise. Evaluation of the effectiveness of the proposed methodology can be done for a larger data set.

VIII. REFERENCES

• Yuanyuan Wu, Eric W.T. Ngai, Pengkun Wu, Chong Wu, Fake online reviews: Literature review, synthesis, and directions for future research, Decision Support Systems, Volume 132, 2020, 113280, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2020.113280.

• W. Liu, J. He, S. Han, F. Cai, Z. Yang and N. Zhu, "A Method for the Detection of Fake Reviews Based on Temporal Features of Reviews and Comments," in IEEE Engineering Management Review, vol. 47, no. 4, pp. 67-79, 1 Fourth quarter,Dec. 2019, doi: 10.1109/EMR.2019.2928964.

• R. Mohawesh. "Fake Reviews Detection: A Survey," in IEEE Access, vol. 9, pp. 65771-65802, 2021, doi: 10.1109/ACCESS.2021.3075573.

• Ahmed, H., Traore, I., Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore, I., Woungang, I., Awad, A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science (), vol 10618. Springer, Cham. https://doi.org/10.1007/978-3-319-69155-8_9

• Atefeh Heydari, Mohammadali Tavakoli, Naomie Salim, Detection of fake opinions using time series, Expert Systems with Applications, Volume 58, 2016, Pages 83-92, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2016.03.020.

• Paul, H., Nikolaev, A. Fake review detection on online E-commerce platforms: a systematic literature review. Data Min Knowl Disc 35, 1830–1881 (2021). https://doi.org/10.1007/s10618-021-00772-6

• Deng, X., Chen, R. (2014). Sentiment Analysis Based Online Restaurants Fake Reviews Hype Detection. In: Han, W., Huang, Z., Hu, C., Zhang, H., Guo, L. (eds) Web Technologies and Applications. APWeb 2014. Lecture Notes in Computer Science, vol 8710. Springer, Cham. https://doi.org/10.1007/978-3-319-11119-3_1

• P. Rathore, J. Soni, N. Prabakar, M. Palaniswami and P. Santi, "Identifying Groups of Fake Reviewers Using a Semi Supervised Approach," in IEEE Transactions on Computational Social Systems, vol. 8, no. 6, pp. 1369-1378, Dec. 2021, doi: 10.1109/TCSS.2021.3085406.

• Khan, H., Asghar, M.U., Asghar, M.Z., Srivastava, G., Maddikunta, P.K.R., Gadekallu, T.R. (2021). Fake Review Classification Using Supervised Machine Learning. In: et al. Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science (), vol 12664. Springer, Cham. https://doi.org/10.1007/978-3-030-68799-1_19

• Li, Y., Wang, F., Zhang, S. et al. Detection of Fake Reviews Using Group Model. Mobile Netw Appl 26, 91–103 (2021). https://doi.org/10.1007/s11036-020-01688-z