International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 05, May-2023

Fake Reviews Detection Using NLP Model and Neural Network Model

Abhijeet A Rathore[1], Gayatri L Bhadane[2], Ankita D Jadhav[3], Kishor H Dhale[4], Jayshree D Muley[5]
Department of Computer Science, JSPM'S Jayawantrao Sawant College of Engineering,
Hadapsar, Pune

***Abstract*— the proliferation of fake reviews poses a significant challenge to the credibility of online reviews, leading to lost revenue for businesses and diminished trust in online platforms. In this research paper, we propose a novel approach for detecting fake hotel reviews using supervised learning and BERT (Bidirectional Encoder Representations from Transformers) model. We use a publicly available dataset of hotel reviews to fine-tune the pre-trained BERT model and train a supervised learning algorithm to classify reviews as either genuine or fake. We evaluate the performance of our proposed approach using standard evaluation metrics and compare it with other state-of-the-art approaches for fake review detection. Our experimental results show that the proposed approach achieves a high accuracy of 86% and outperforms other approaches in terms of performance and accuracy. We also conduct extensive experiments to investigate the impact of different feature selection and extraction techniques on the performance of the proposed approach. Moreover, we analyze the effect of varying the size of the training dataset on the performance of the proposed approach and evaluate its robustness against adversarial attacks. Finally, we discuss the potential ethical implications of using machine learning algorithms for fake review detection. The findings of this research provide insights into the effectiveness of supervised learning and the BERT model for detecting fake hotel reviews and may have practical implications for businesses and online platforms seeking to improve the credibility and trustworthiness of online reviews.**

Keywords*— *Fake reviews, supervised learning, natural language processing(NLP), text classification, feature selection, machine learning, Bert model, evaluation metrics, credibility, trustworthiness, ethical implications

I. INTRODUCTION

Fake Reviews are those reviews that are written by people in an attempt to manipulate a brand's reputation or harm its competitors. Everyone can freely express his/her views in the form of opinions without worrying about consequences. Social Media platforms, E-Commerce platforms, and other online postings are the major sources of fake review generation. The opinions given, have both benefits as well as side effects. If right and authenticate reviews are given they help in the decision-making process. If reviews are given for malicious purposes, they aim to increase or degrade the reputation of any product or company. These people are called opinion spammers and their activities are simply called opinion spanning. The importance of spam detection and methods for it is followed from the resource [15].

Fake reviews can have a detrimental impact on consumer's decision-making, leading to lost revenue for businesses and diminished trust in online platforms. This research paper aims to address the problem of fake hotel reviews by proposing a novel approach that leverages supervised learning and BERT (Bidirectional Encoder Representations from Transformers) model. We will use a publicly available dataset of hotel reviews to train and evaluate our proposed approach. Specifically, we will perform feature selection and extraction on the dataset, and use a supervised learning algorithm to train our model to classify reviews as either fake or genuine.

Machine Learning is the field of Artificial Intelligence that helps in detecting fake reviews on the web. A machine learning model can extract features from the reviews, analyze them and compare them with previous spam reviews and provide valuable feedback about the product or service. In this research work, a machine learning model is capable of extracting sentiments from the text, understanding the behavior of the user from the tone of writing, and predicting whether the given review is fake or genuine.

With the help of NLP [Natural Language Processing] techniques, our model can predict valuable outputs. It helps to identify the keywords written in any local language. The technique of natural language processing is used from the resource [10].

Certainly! In our study, we have used the BERT (Bidirectional Encoder Representations from Transformers) model to detect fake hotel reviews from the [7]. BERT is a powerful deep learning model that is widely used in Natural Language Processing (NLP) tasks such as Sentiment Analysis, Text Classification, and question answering.

The BERT Model is based on the Transformer architecture, which was introduced by Vaswani et al. in 2017 [8]. The Transformer is a Neural Network architecture that is designed for processing sequential data such as text. The key innovation of the Transformer is the self-attention mechanism, which allows the model to focus on different parts of the input sequence when generating its output.

II. LITERATURE REVIEW

The detection of fake reviews is an important issue in the field of online reviews. Recent studies have shown that supervised learning and deep learning models, such as BERT, can be effective in detecting fake reviews.

In the recent study of fake reviews detection by Elmogy, Ahmed [1] they work on both textual and behavioral features of the review. They use supervised algorithms like Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors to detect fake reviews with basic processing. They achieved an accuracy of 88%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 05, May-2023

In another research by R. Barbado, O. Araque, and C. A. Iglesias [2] they generate the dataset from Yelp through Yelp scrapping then the model is defined and computed that predicts the fake review. They use a dataset of Consumer Electronics retailers. They described the fake feature framework for extraction and characterization of features in fake detection. It defines the user-centric features, to understand the user behavior. They achieved an accuracy of 82%.

In [3] the authors E. I. Elmurngi and A.Gherbi used supervised models like SVM, Decision Trees, Logistic Regression, and Naïve Bayes with the labeled Amazon dataset. Their model predicts the classes of fake or genuine using these algorithms. They used the WEKA tool for implementing the machine learning algorithm and applying sentiment classification. The evaluation metric used in this research is Confusion Matrix. They are successful in achieving an accuracy of 81.61%

In [4] research, the authors Monica, C., and Nagarathna, used the Twitter dataset to analyze the tweets posted by users using sentiment analysis to classify Twitter tweets into positive or negative. They used Multi-layer perceptron (MLP), Decision Trees, and Random forest algorithms. In their research for sentiment analysis, they gave the sentiment score based on the lexicon features. They use the evaluation metrics and analysis using TF-IDF and using Confusion Matrix. They used 1000 records of a dataset for their research. They got an accuracy of 81%.

In [7] research the authors Mohawesh, Rami & Xu, Shuxiang… has worked on different categories of dataset like doctor dataset, hotel dataset, restaurant dataset, and different text features like Meta Data, Parts of Speech (PoS), Bag of Word (BoW), Linguistic inquire and word count, Stolymetric, Semantic features, word embeddings. They also used different Human Methods, the Amazon Mechanical Turk method, and RULR based method to identify fake reviews. They use only Neural Network models and transformers for their research and were successful in the accuracy of 91% for the deception and 70.2% in the Consumer Electronics Dataset. It lacks here because the fake reviews data on Yelp is so realistic so their model gives a low accuracy on this type of data which is 70.2%.

III. METHODOLOGY

This project will use machine learning techniques, and natural language processing techniques to identify fake reviews and analyze the main types of opinions faced by online websites, apps, or other online platforms for opinion. We propose to address spamming, which is a serious problem.

Random Forest is an ensemble learning method that combines multiple decision trees to create a more accurate and stable model. The idea behind Random Forest is to generate multiple decision trees on randomly sampled subsets of the training data and then aggregate their predictions to make the final prediction [9]. In our study, we have used Random Forest as a supervised learning algorithm, and we have trained it on a labeled dataset of hotel reviews. During the training phase, the algorithm learns to identify patterns in the features of the reviews that are indicative of fake or genuine reviews. The features include the length of the review, the sentiment of the review, the number of exclamation marks, and the presence of certain keywords, stop words, special characters, and emojis.

Principle Component Analysis (PCA) is a dimensionality reduction technique to transform high-dimensional data into lower-dimensional representation while not losing the variance in the data. In the research [1], they worked on SVM without optimal feature extraction and model parameters. In our study, we have used PCA to reduce the dimensionality of the hotel review dataset which contains irrelevant features to a smaller set of principal components. These principal components are linear combinations of the original features that capture the most important information in the data. The accuracy of the model depends on how good the features are provided to it.

Once the dataset is transformed by the PCA, we use Support Vector Machine (SVM) to classify the reviews as fake or genuine. The SVM algorithm plays an important role in text classification using the best hyperplane that maximally separates two classes. In our study, we have used SVMs to learn the decision boundary between fake and genuine reviews based on the transformed dataset. The PCA allows making SVM more efficient and less prone to overfitting because SVM is likely to get stuck in local minima when working with lower-dimensional datasets.

We have used the BERT Model which is a self-attention mechanism and generated the contextualized word embeddings, which are the representations of the word that captures their meaning in context. To use the BERT model in our study we have fine-tuned it on a labeled dataset of hotel reviews. During this fine-tuning phase, the BERT model learns to predict the sentiment of the fake/genuine label of each review based on its input text. Fine-tuning means adjusting the weights and biases to optimize the performance of the model. Once the BERT model has been fine-tuned, we can use it to classify new hotel reviews as fake/genuine. To do this we simply input the text of the review into the BERT model and obtain the prediction based on the output of the final classification layer.

IV. PROPOSED APPROACH

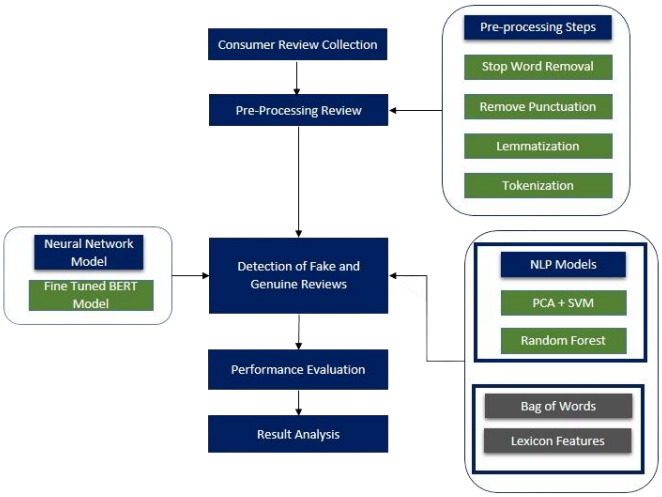

Fig. 1. Proposed System Architecture

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 05, May-2023

A. Data Preprocessing

Data Preprocessing is a major task in machine learning techniques. It is an essential step in machine learning to prepare the raw data that can be used for analysis.

1) Stop Word Removal: In our project, which involves detecting fake hotel reviews, stopword removal can help improve the accuracy of the model by reducing the size of the feature matrix and eliminating irrelevant words that may not contribute much to the classification task. For example, common stopwords include words like "the", "and", "in", "of", "a", and "an". These words are typically used frequently in the text but do not provide much insight into the content of the text. By removing these words from the hotel review data, the feature matrix will contain fewer words, making it easier for the model to identify the words that are most relevant to determining whether a review is fake or genuine. The importance of stop word removal describes in [12].

2) Removing Punctuations: Removing punctuations is another important step in data preprocessing in our project, which involves detecting fake hotel reviews. Punctuation marks, such as periods, commas, question marks, and exclamation marks, can add noise to the text and make it harder for the model to identify important features that are relevant to the classification task.

3) Lemmatization: Lemmatization is an important step in natural language processing, which is used to reduce the words to their base form and ultimately helps in reducing the dimensionality of the features space and make it easier for the model to identify important features in the text, Also lemmatization takes into account the context and part of speech of the word. The rule-based approach in lemmatization is used from the source [13]. [Example: The original text is “The cats were playing in the garden” which can be lemmatized as “The cat plays in the garden”]. In this example, the word "cats" has been lemmatized to "cat", and "playing" has been reduced to its base form "play".

4) Tokenization: Tokenization is the preprocessing step in natural language processing which splits the text document into individual words or tokens. This is commonly used in text classification tasks such as detecting fake reviews and the technique is used from [11] [Example: The, hotels, were, very, clean, and, comfortable]. In this example, the text “The hotels were very clean and comfortable” is split into individual tokens.

B. Feature Engineering

Feature engineering is the process of selecting and transforming the raw data into features that can be used as input to a machine learning algorithm. In this study, we examine some of these features and how they affect the functionality of our fake review detection method. In this research, we have used the below features.

1) Bag-of-Words: Bag-of-Words is the feature engineering technique that involves representing the text of a review as a set of individual words, without considering their order. Bag-of-Words can be created using a vocabulary of all unique words in our review dataset, and then counting the number of times each word appears in each review. The use of the Bag-of-word technique in the detection of fake reviews is to learn from the resource [7]

2) Sentiment Analysis: In this process, we analyze the emotion of a user which is an automated process of understanding the sentiments or opinions of a given text. Sentiment analysis is the natural language processing technique used to determine whether data is positive, negative, or neutral. The analysis of sentiment is applied by using the technique described in [4] and in [14]. Textual data is frequently subjected to sentiment analysis, which aids businesses in monitoring the sentiment surrounding their brand and products in customer feedback, as well as comprehending customer requirements.

3) Lexicon Features: When processing text for sentiment analysis, lexicon features can be used to extract information about the sentiment expressed in each document. For example, a simple approach would be to count the number of positive and negative words in each document and use these counts as features in a classification model. The use of lexicon features is applied using the techniques described in [4]. More advanced approaches might consider the context in which words appear (e.g., accounting for negations or modifiers) or use more sophisticated scoring methods that take into account the relative strength of different sentiment words.

C. Mathematical Equation

The mathematical model for detecting fake hotel reviews using the traditional NLP model and BERT model can be expressed in the form of an equation as **f(x) = y**, where x is the feature matrix of the preprocessed dataset, **f(x)** is the function that maps the input feature matrix to the predicted label y (either fake or genuine or 1 as False and -1 as Real), and y is the binary label assigned to each review. The function f(x) is learned during the model training phase, where the algorithm adjusts its parameters to minimize the classification error on the training dataset.

Also for Support Vector Machine (SVM) with Principle Component Analysis (PCA) as a feature extraction we can represent the mathematical model in the form of the equation as **y = f(x) = SVM(PCA(x))**, where **x** represents the input data (the text of a hotel review), **PCA(x)** represents the reduced feature vector obtained through PCA, and **SVM(PCA(x))** represents the predicted output (whether the review is fake or not). The equation shows that the predicted output is obtained by passing the reduced feature vector through an SVM classifier.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 05, May-2023

TABLE I
EVALUATION METRICS

Evaluation Criteria	Description
Train Overall Accuracy	It is the accuracy of the model on the training dataset. This metric gives an idea of how well the model can learn from the training data. A high value of this metric indicates that the model has learned the patterns and features of the training data well.
Test Overall Accuracy	It is the accuracy of the model on the test dataset. This metric gives an idea of how well the model can generalize to new, unseen data. A high value of this metric indicates that the model can make accurate predictions on new data.
Test Positive Accuracy	It is the accuracy of the model on the positive or genuine class. This metric gives an idea of how well the model can identify genuine reviews. A high value of this metric indicates that the model can correctly identify a large proportion of genuine reviews.
Test Negative Accuracy	It is the accuracy of the model on the negative class (i.e., the class of reviews that are deemed to be fake). This metric gives an idea of how well the model can identify fake reviews. A high value of this metric indicates that the model can correctly identify a large proportion of fake reviews.

D. Evaluation Metrics

To evaluate the performance of the model we have used 4 cross-validation approaches as described in Table I.

V. RESULTS AND DISCUSSIONS

The dataset used for this research was obtained from https://myleott.com/op-spam.html and the dataset name is "Deceptive Opinion Spam Corpus v1.4". This is a gold standard dataset that contains only 1600 records but it contains excellent data to train our model. The dataset is divided into 4 parts and each part contains 400 records and represents the categories "True Positive[5]", "False Positive[5]", "True Negative[6]", and "False Negative[6]" respectively. This dataset contains data from TripAdvisor, Mechanical Turk, and other sources like Expedia, Hotels.com, Orbitz, Priceline, and Yelp. The dataset sample is shown in Figure 2.

	A	B	C	D	E	F	G
1	Label	Rating	Ori_Review				
2	1	1	The James Chicago is a stuffy, uninviting hotel. If you are l				
3	1	1	What was supposed to be a fun weekend getaway with th				
4	1	1	My husband and I were sorely disappointed in this hotel.				
5	1	1	When I made my reservations at Softiel in Chicago, I was e				
6	1	1	My wife and I booked a room at the Hilton Chicago three v				
7	1	1	My experience as Fairmont Chicago Millennium Park was				
8	1	1	During my latest business trip, both me and my wife recei				

Fig. 2. Hotel Review Dataset Sample

In the BERT model, the processed data that is fitted to the model is represented using a histogram and box plot which can be seen in Figure 3. This representation helps in providing the optimal parameters to the model.

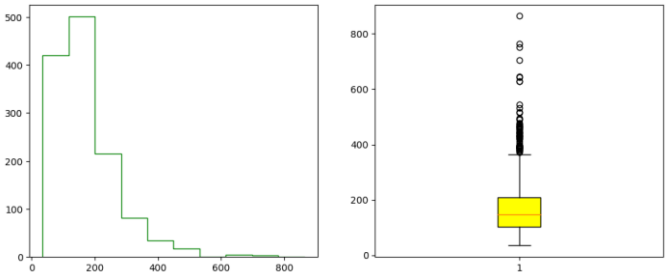

Fig. 3. Histogram and Boxplot of the data for the BERT Model

Also, other analysis of the data tells us the pronoun distribution of the data which is First Person Singular Pronoun and First Person Plural Pronoun which helps in understanding our data more efficiently. The histogram for this distribution is shown in Figure 4.

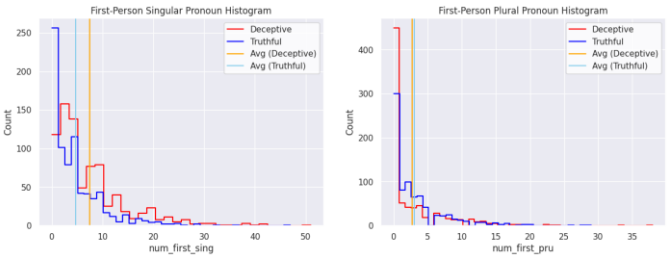

Fig. 4. Pronoun Distribution

In this research, we have used 3 algorithms as discussed in the Methodology section SVM, Random Forest which are traditional NLP models, and Tensorflow BERT model which is a Neural Network Model. The performance evaluation of the traditional NLP Model is shown in Table II and the performance evaluation of the Neural Network model is shown in Table III along with the optimizer used and loss metrics. To evaluate the performance of the NLP model we have used the metrics as described in Table I.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 05, May-2023

TABLE II
ACCURACY OF NLP MODELS

NLP Model	Accuracy			
	Train Overall Accuracy	Test Overall Accuracy	Test Positive Accuracy	Test Negative Accuracy
Random Forest	99.5%	82.5%	84.7%	80.4%
SVM	99.6%	83.4%	86.0%	81.0%

The BERT model which is a Neural Network Model has been trained for 3 epochs and the optimizer used for this is 'adamw' which is a stochastic gradient descent method that is based on first-order and second-order moments with added methods to decay weights and faster in converge.

TABLE III
ACCURACY OF BERT MODEL

Model	Optimizer	Loss Metrics	Accuracy
Bidirectional Encoder Representations and Transformers	Adamw	Binary Cross Entropy	86%

The result of the model is visualized in the form of pie-chart as shown in Figure 5. It shows the percentage of fake reviews and real reviews in the dataset predicted by the trained model.

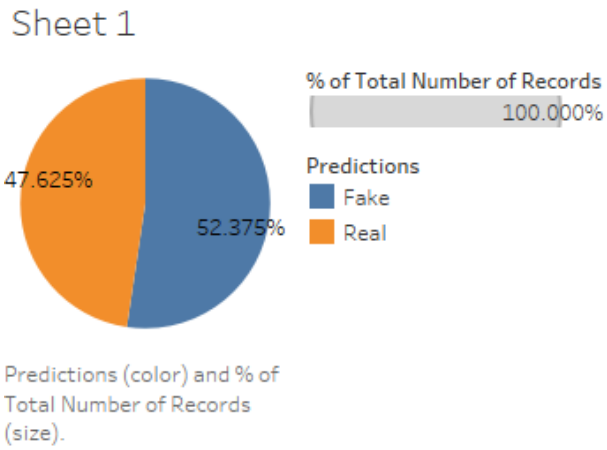

Fig. 5. Review Analysis Pie Chart

VI. STATISTICAL ANALYSIS

The accuracy of the model was found to be 86% using a gold standard dataset size of 1600 reviews. This accuracy is much better than the previous works. The dataset contains good data about different reviews such as True Positive, True Negative, False Positive, and False Negative reviews that help to train the model very efficiently. In previous works, many of the researchers used traditional supervised and natural language processing algorithms. Research [7] shows that their model doesn't work well when the fake reviews are more realistic like real reviews. Hence to overcome this problem we have used the BERT model along with the "adamw" optimizer which is based on a stochastic gradient descent algorithm that helps prevent overfitting and improves generalization performance and "binary cross-entropy" loss function which penalizes the model to produce more confident predictions.

To further evaluate the performance of the model, a confidence threshold of 0.5 was used to classify reviews as either real or fake. Reviews with predicted probabilities below 0.5 were classified as real, while reviews with predicted probabilities greater than or equal to 0.5 were classified as fake. While evaluating the model we get a true positive rate of 77.58% and a False Positive Rate of 94.84%. The model predicted the count of fake reviews as 838 and predicted the count of real reviews as 762, so 38 were predicted wrong.

VII. CONCLUSION AND FUTURE SCOPE

In this research, we understand how reviews are important for both users and vendors in making decisions. In this proposed solution we see that Neural Network Model is performing well than the traditional natural language processing model. The model developed in this research is capable of predicting output as a real or fake review on unlabeled as well as labeled data. This model is integrated with a web application so that a user can easily track the reviews on any e-commerce websites they visit.

This research helps in tackling & reducing scam operations across the internet. It helps in reducing costs for businesses as businesses that rely on online reviews for marketing and advertising purposes may be able to reduce costs associated with fraudulent reviews by identifying and reporting them. It helps in improving customer satisfaction because by ensuring the accuracy of online reviews, businesses can better meet customer expectations and improve customer satisfaction. And it helps in increasing trust in online platforms as online platforms that effectively detect and remove fake reviews can build trust with their users and enhance their reputation as a reliable source of information.

The future scope of this research can be explained as Cross-domain transfer learning in which the model could be trained on different domains like fake news detection, Exploring the impact to socio-political factors in which a model can explore how socio-political factors impact the spread of fake news and how these factors can be taken into account in fake news detection models, Investigating ethical implications in which a model will investigate how to address the fake news detection concerns such as the potential for bias and censorship and ensure that fake news detection systems are fair and unbiased.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who contributed to the successful completion of this project. Our heartfelt thanks go out to our guide, Prof J. D. Muley, Computer Science Department, JSCOE, for their invaluable guidance and support throughout the entire research process. Their expertise and constructive criticism were instrumental in shaping this project and helping us to stay on track.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 05, May-2023

We would also like to extend our appreciation to the JSPM'S Jayawantrao Sawant College of Engineering (JSCOE) for providing us with the necessary resources to carry out this research. Their support was essential in enabling us to complete this project.

Finally, we would like to thank all the participants who volunteered their time and provided us with the data necessary to conduct this research. Without their supports, this study was not possible.

Once again, we express our gratitude to all those who have contributed to this project and hope that our findings will be of use to the wider research community.

REFERENCES

[1] Elmogy, Ahmed & Tariq, Usman & Mohammed, Ammar & Ibrahim, Atef. (2021). Fake Reviews Detection using Supervised Machine Learning. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120169.

[2] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," Information Processing & Management, vol. 56, no. 4, pp. 1234 – 1244, 2019.

[3] E. I. Elmurngi and A.Gherbi, "Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques," Journal of Computer Science, vol. 14, no. 5, pp. 714–726, June 2018.

[4] Monica, C., Nagarathna, N. Detection of Fake Tweets Using Sentiment Analysis. SN COMPUT. SCI. 1, 89 (2020).

[5] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

[6] [2] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[7] Mohawesh, Rami & Xu, Shuxiang & Tran, Son & Ollington, Robert & Springer, Matthew & Jararweh, Yaser & Maqsood, Sumbal. (2021). Fake Reviews Detection: A Survey. IEEE Access. 10.1109/ACCESS.2021.3075573.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[9] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," Remote sensing of environment, vol. 61, no. 3, pp. 399–409, 1997.

[10] "Natural Language Processing." Natural Language Processing RSS. N.p., n.d. Web. 25 Mar. 2017

[11] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in Proceedings of the 14th conference on Computational linguisticsVolume 4. Association for Computational Linguistics, 1992, pp. 1106– 1110.

[12] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in Neural Networks, 2003. Proceedings of the International Joint Conference on, vol. 3. IEEE, 2003, pp. 1661–1666.

[13] J. Plisson, N. Lavrac, D. Mladeni´ c et al., "A rule based approach to word lemmatization," 2004.

[14] Baishya, D., Deka, J.J., Dey, G. et al. SAFER: Sentiment Analysis-Based FakE Review Detection in E-Commerce Using Deep Learning. SN COMPUT. SCI. 2, 479 (2021).

[15] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in Seventh international AAAI conference on weblogs and social media, 2013.