International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Implementation of Multiview Deep Learning Algorithm using LSTM-CNN for Fraudulent Review Detection

Priyanka Gupta[1]
School of Computer Science and Engineering
Lovely Professional University Punjab, India
priyanka.21789@lpu.co.in

Abhishikth Xavier Kapudasi[1]
School of Computer Science and Engineering
Lovely professional University Punjab, India
abhikapudasi1407@gmail.com

Rakesh Eemani[1]
School of Computer Science and Engineering
Lovely Professional University Punjab, India
rakesh70952@gmail.com

Kaza Chandu Koundinya[1]
School of Computer Science and Engineering
Lovely Professional University Punjab, India
koundinyakaza@gmail.com

Sujith Dugyala[2]
School of Computer Science and Engineering
Lovely Professional University Punjab, India
sujithdugyala25@gmail.com

***Abstract* - In this project, we explored the task of detecting fraudulent reviews in text data using various machine learning algorithms. The dataset consisted of reviews categorized into two labels: 'CG' (genuine reviews) and 'OR' (potentially fraudulent reviews). We began by preprocessing the text data, which involved tasks such as removing punctuation, stop words, and applying stemming or lemmatization techniques. After preprocessing, we utilized different machine learning models including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machines (SVM), Multinomial Naive Bayes and LSTM-CNN model to classify the reviews. Overall, our study showcases the effectiveness of machine learning, Deep learning algorithms in identifying fraudulent reviews from textual data. The findings provide valuable insights for applications in e-commerce platforms and online review systems to enhance the reliability and trustworthiness of user-generated content.**

Keywords: Fraudulent reviews, Text classification, Machine Learning, Preprocessing, Logistic Regression, Support Vector Machines, Evaluation metrics.

I. INTRODUCTION

In today's world where everything is online, reviews really matter when people are deciding what to buy or where to go. Reviews give us important information when we are looking for something. But fraudulent reviews are a big problem because they make it hard to trust reviews.

In this paper, we're trying to find fraudulent reviews using advanced machine learning techniques. Our data has a bunch of reviews labelled as either "CG" for real reviews or "OR" for possibly fraudulent reviews.

First, we did some cleaning of the text data. This included taking out punctuation, common words, and standardizing the text using things like stemming.

After getting the data ready, we looked at different machine learning algorithms that could classify the reviews as real or fake. We tried Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Naive Bayes, and LSTM-CNN. We wanted to see which worked best at detecting fakes. To figure that out, we evaluated the models in many ways using things like precision, recall, accuracy, and confusion matrices. These helped us understand how well each model correctly labelled the reviews. The LSTM-CNN model did the best, getting an accuracy of around 92%, SVM with 88%, Logistic Regression also perform well at 86% accuracy. Decision Trees, Random Forests, and Naive Bayes got around 73%, 84%, and 84% accuracy. In the end, this project shows that machine learning, Deep learning can help identify fraudulent reviews in text data. Finding fakes is important so reviews can be trusted and help people make better buying choices in today's digital world.

II. LITERATURE REVIEW

Asaad, Allami, and Ali (2023) introduced a technique for classifying and identifying fraudulent reviews in online platforms using machine learning methodologies.[2] Their study focused on the detection of fraudulent reviews in the context of hotel services, utilizing TFIDF techniques and three machine learning algorithms: Xgboost, support vector classifier, and stochastic gradient descent. The proposed model was evaluated on balanced and imbalanced datasets, demonstrating promise in enhancing the credibility of online reviews.

[3] Thirunavukkarasu et al. (2023) presented an approach based on supervised learning for identifying fraudulent reviews in online datasets. Their study utilized machine learning methods such as KNN, Naive Bayes, and logistic

regression to distinguish between false and real reviews. By leveraging labelled datasets and traditional machine learning algorithms, the authors aimed to combat the spread of false reviews that deceive online shoppers and impact the reputation of businesses.

Other than these there are many studies which also specify the accuracy scores of the models which they have used. Some of those studies are listed in the table -1.

S. No	Paper Number	Authors	Method	Accuracy Score
1	1	A. M. Elmogy, U. Tariq,	Supervised Machine Learning	82.40%
2	6	Ahsan, M.N.I., Nahian, T., Kafi, A.A., Hossain, M.d.I., Shah, F.M.	Ensemble Learning	Above 95%
3	8	Bali, A.P.S., Fernandes	Machine Learning (Gradient Boosting)	88%
4	10	M. S. Jacob, S. Rajendran, V. M. Mario, K. T. Sai, D. Logesh	Supervised Machine Learning (Logistic Regression, Linear Regression, CNN, RNN)	Optimal Accuracy.

Table-1: Similar studies

[4] A. Moqueem, F. Moqueem, Reddy, Jayanth, and Brahma (2023) proposed various machine learning models for detecting fraudulent reviews in online shopping platforms. Their work emphasized the importance of online reviews in influencing purchasing decisions and highlighted the prevalence of fraudulent reviews in misleading consumers. By comparing and analysing different machine learning algorithms, the authors aimed to contribute to reducing and checking the spread of fraudulent reviews.

[5] In their study, Mohawesh and colleagues (2021) undertook a thorough examination of methods for identifying fraudulent reviews, delving into various datasets, techniques for extracting features, and strategies for detection. Their research underscored the critical role of fraudulent review detection in upholding the trustworthiness of online platforms and safeguarding the interests of consumers. Moreover, they carried out benchmark assessments to gauge how well different neural network models and transformers performed, shedding light on potential avenues for future investigation in this field.

[7] Liu, He, Han, Cai, Yang, and Zhu (2019) proposed a method for detecting fraudulent reviews based on temporal features of reviews and comments. Their study focused on analysing review records associated with products, utilizing an isolation forest algorithm to detect outlier reviews. By investigating the differences between patterns of product reviews, the authors aimed to provide a new perspective on outlier review detection and enhance the effectiveness of fraudulent review detection methods.

[9] Bhat, Jayalakshmi, Mallegowda, and Geetha (2024) presented a comprehensive approach to detecting fraudulent reviews using both supervised and unsupervised learning techniques. Their study leveraged classic machine learning algorithms, deep learning techniques such as RNN and attention networks, as well as state-of-the-art models like BERT and GPT. By comparing the performance of different models and interpreting their results, the authors aimed to provide practical solutions for combating fraudulent reviews and ensuring the integrity of online platforms.

After reviewing the literature on fraudulent review detection, it is evident that researchers have employed various methodologies to address this pressing issue in online platforms. Studies have emphasized the importance of accurately identifying fraudulent reviews to maintain the credibility of e-commerce platforms and protect consumer interests. Approaches range from traditional machine learning algorithms to advanced deep learning techniques, each offering unique insights into the detection process. Benchmark studies have provided valuable comparisons of different models, highlighting their strengths and limitations. Overall, this body of research contributes to the ongoing efforts to combat fraudulent reviews and offers avenues for further exploration in the field of online review integrity.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

III. ABOUT THE DATA

Size: The dataset contains 40,432 entries.

Columns:

- category: Indicates the category of the product or service associated with each review.
- rating: Represents the rating given to each review, ranging from 1 to 5.
- label: Indicates a label associated with each review. (e.g., "CG" or "OR")
- text_: Contains the text data of the reviews.

The dataset is pre-processed for analysis, including text cleaning, preprocessing, and feature extraction. The ratings and labels provide additional context for the reviews, which can be useful for sentiment analysis or classification tasks.

The dataset's primary purpose appears to be the development of a model for fraudulent reviews detection, utilizing various machine learning algorithms to classify reviews based on their authenticity or sentiment.

IV. DATA EXPLORATION

Data exploration is important as we need to know about the data set and the relations between the data but in this project, we will be going to do a basic data exploration which includes the following:

a. knowing about the data set by printing it.

	category	rating	label	text_
0	Home_and_Kitchen_5	5.0	CG	Love this! Well made, sturdy, and very comfor...
1	Home_and_Kitchen_5	5.0	CG	love it, a great upgrade from the original. I...
2	Home_and_Kitchen_5	5.0	CG	This pillow saved my back. I love the look and...
3	Home_and_Kitchen_5	1.0	CG	Missing information on how to use it, but it i...
4	Home_and_Kitchen_5	5.0	CG	Very nice set. Good quality. We have had the s...

Table-2: The raw data present in the data set

b. plotting pie chart for knowing the proportion of each rating.

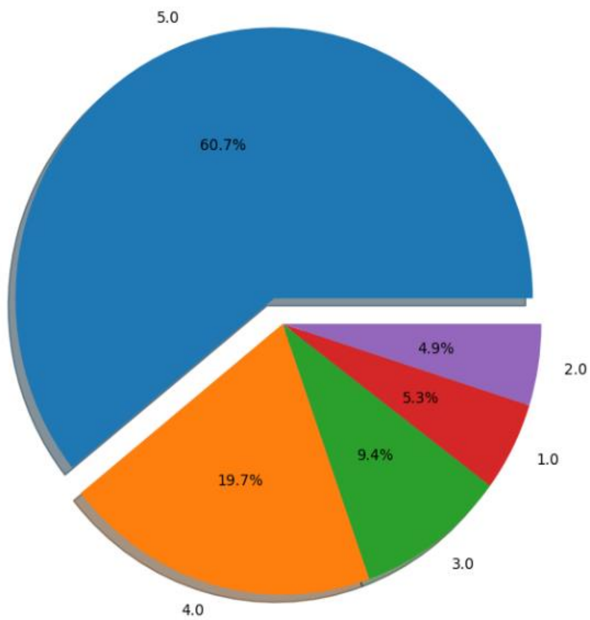

Figure -1: The division of ratings in the data

We see that 5.0 rating is dominating the chart.

c. Plotting line graph for checking the relation between Length of the review and Index.

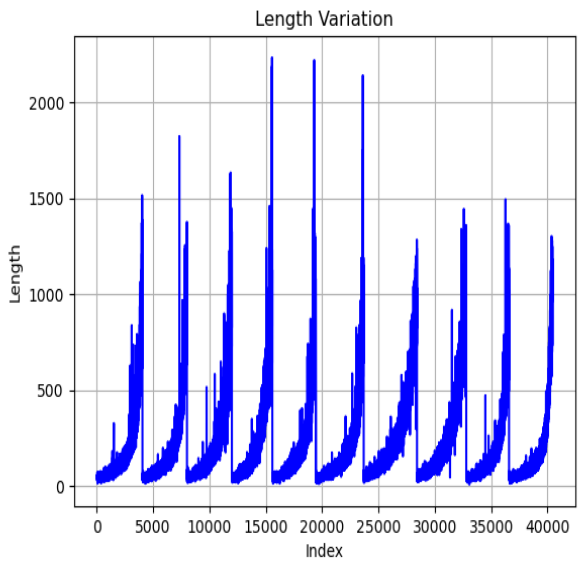

Figure-2: Line graph for Length and Index of the data

We see that near index value 15000 to 25000 the length of the words is high.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

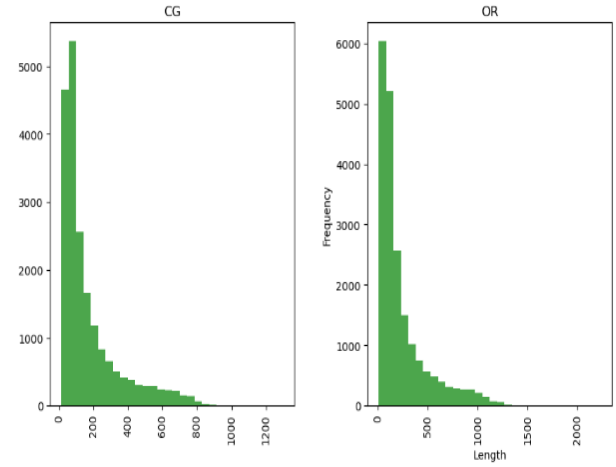

Figure-3a: Bar graph for Distribution of length by label in CG and OR

We observe from Figure-3a that at 0 length the frequency is highest in OR

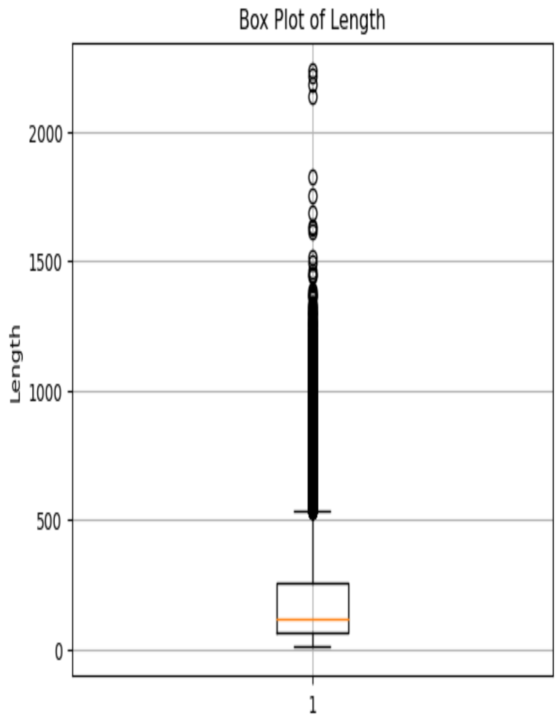

Figure-3b: Box Plot for the Length of the review

These was no use of the box plot because length of reviews was having different lent. Thats the reason there are many outliers in the boxplot.

V. PROPOSED MODEL OVERVIEW

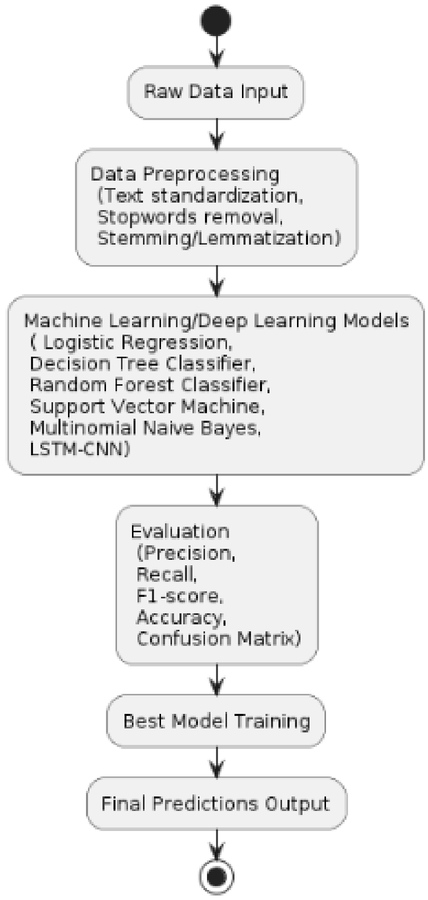

Figure -4: Proposed Model

From Figure-4 We can see the basic structure of the proposed model. The flow chart goes as below:

Input: This initial step represents the input which is given to the model so that it can detect fraudulent reviews.

1. Data Preprocessing:
   a. Text Standardization: The data which is given in raw form is converted into consistent form, such as lowercasing all letters.
   b. Stop words Removal: Then we have removed all the stop words such as “the”, “and”, “is” from the text data, as they don’t carry significant meaning for analysis.
   c. Stemming/ Lemmatization: In this step we have converted the word into their root form by removing the suffixes.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

2. Machine Learning Models:
   a. Logistic Regression: This algorithm used for binary classification tasks. We used it in our model, and we received an accuracy of 86.11%.
   b. Decision Tree Classifier: This Model is a tree type classifier which splits the data according to the features. By using this model, we received 72.79 % of accuracy.
   c. Random Forest Classifier: This classifier ensembles number of random forest tress together to predict the overall output. As we have used this model in our project, we have gained 83.85% of accuracy.
   d. Support Vector Machines (SVM): A supervised algorithm, which is generally used for classification, after using this model we have received the highest accuracy of 87.99%.
   e. Multinomial Naive Bayes: This classifier uses probability as its base for predictions. It uses Bayes' theorem for getting probability score. By using this model, we have received 84.43% of accuracy.
   f. LSTM-CNN: This combination of Deep learning model performed well when compared to other models. Overall accuracy of the model was 92% which was better than any other model.

We did use five kinds of machine learning algorithm with one kind of Deep learning algorithm to know which model performs better in the field. From the output of the models, we can easily tell that LSTM-CNN has performed better than any other model in our project, as LSTM-CNN model uses a combination of Machine and Deep learning model. It has

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.86	0.84	0.86	0.86
Decision Tree Classifier	0.73	0.73	0.73	0.73
Random Forest Classifier	0.83	0.80	0.83	0.83
Support Vector Machines (SVM)	0.88	0.90	0.88	0.88
Multinomial Naïve Bayes	0.84	0.81	0.84	0.84
LSTM - CNN	0.92	0.94	0.92	0.92

given us the best output.

VI. RESULTS

Evaluation of the model:

To evaluate the models, we have used precision score, Recall score, F1-Score, Accuracy of the model, Confusion Matrix for each model.

Precision and Recall score of the model

Table -3: Accuracy score, Precision, Recall, F1-Score of the models used

Among the machine learning models in the table, Support Vector Machines (SVM) emerged as the top performer, achieving the highest accuracy of 88%. SVM also exhibited excellent Precision and Recall scores, indicating its robustness in correctly classifying instances and minimizing false positives and false negatives. On the other hand, Decision Tree Classifier showed relatively lower performance, with an accuracy of 73% and comparable Precision and Recall scores. The model with the best F1 score, a harmonic mean of Precision and Recall, has a crucial impact on the accuracy score as it represents the balance between precision and recall. In this project, SVM likely contributed significantly to the overall accuracy due to its ability to maintain high precision and recall simultaneously, resulting in a more reliable classification of instances compared to other models.

Other than these Machine Learning Models LSTM-CNN which we have used had given us the best accuracy score of 92%.

Confusion Matrix for the models used:

Using these matrices, we can identify how many inputs were miss classified and how many were correctly classified.

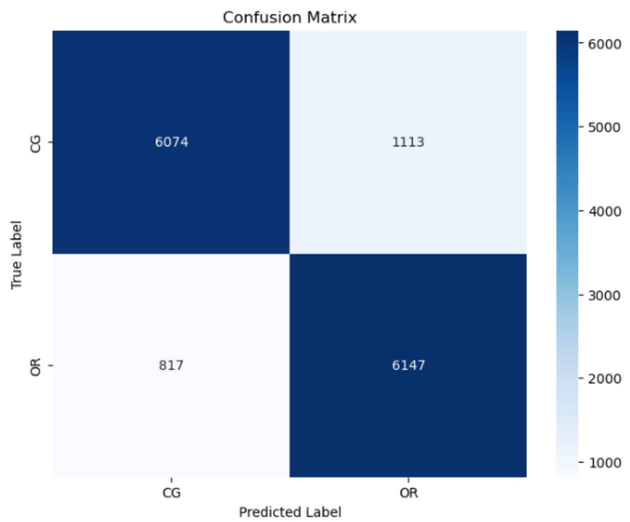

Figure-5: Confusion Matrix for Logistic Regression

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

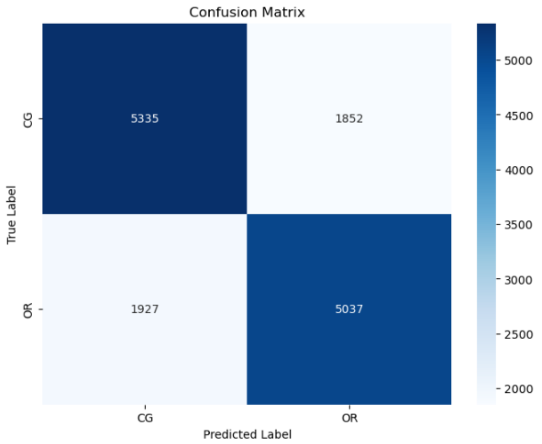

Figure- 6: Confusion Matrix for Decision Tree Classifier

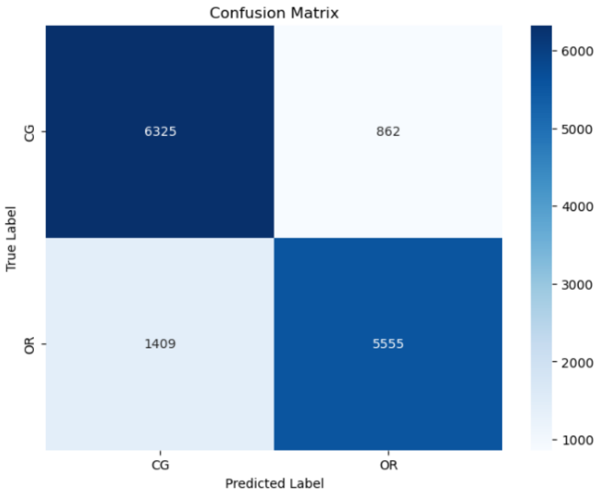

Figure-7: Confusion Matrix for Random Forests Classifier

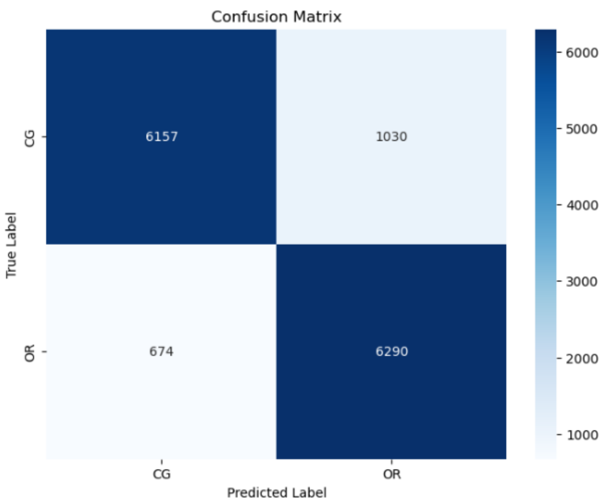

Figure-8: Confusion Matrix for Support Vector machines

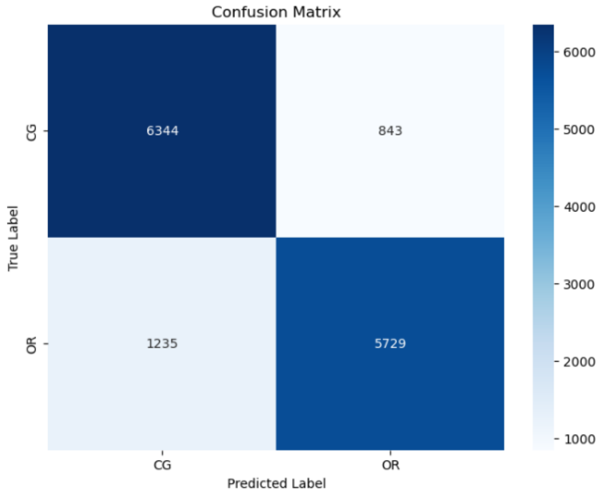

Figure-9: Confusion Matrix for Multinomial Naïve Bayes

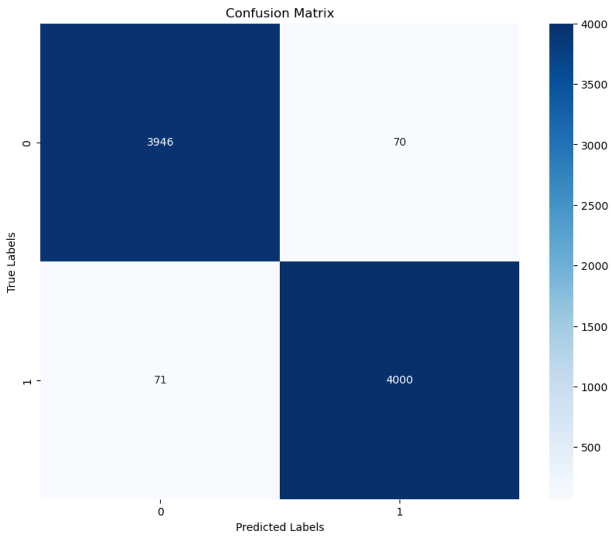

Figure-10: Confusion Matrix for LSTM-CNN

As a result, the LSTM-CNN model had worked better than any other model we happen to perform K-folds method so that the model can perform good with unseen data.

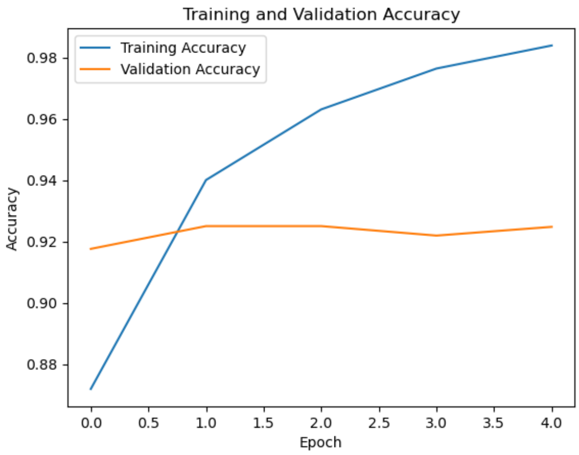

Figure-11: Training and Validation Accuracy graph

From Figure-11 we see that the model had the Validation accuracy of 92% and the Training accuracy of 98%.

VII. CONCLUSION

In conclusion, we have developed a comprehensive model for detecting fraudulent reviews using various machine learning algorithms. Through extensive data preprocessing, including text standardization, stopwords removal, and stemming/lemmatization, we prepared the dataset for analysis. We then employed several machine learning models, including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machines (SVM), Multinomial Naive Bayes, and LSTM-CNN to classify reviews into different categories.

We evaluated the performance of each model using precision, recall, F1-score, and accuracy metrics. The LSTM-CNN model exhibited the highest accuracy, achieving an accuracy score of 92%. However, it's worth noting that the Support Vector Machine also performed well, with an accuracy score of 87.99%.

VIII. FUTURE PERSPECTIVES

Looking forward, there are multiple exciting opportunities for enhancing the fraudulent review detection model and its applications. One such opportunity could be to improve data preprocessing techniques by utilizing more sophisticated natural language processing (NLP) methods. For example, integrating word embeddings or transformer-based models like BERT could help in capturing subtle semantic meanings and enhance the model's comprehension of review texts.

Additionally, improving the model's structure with the latest machine learning and deep learning algorithms could result in superior performance. Strategies such as ensemble learning, attention mechanisms, and neural network designs specifically for text classification tasks could be investigated to increase precision and generalizability.

Also, given the ever-evolving nature of online platforms and user behaviour, it's crucial to continuously retrain and adapt the model. Implementing strategies for incremental learning and active learning can help the model stay effective in identifying changing patterns of fraudulent reviews. Besides improving the model, incorporating specific features from the domain and external data sources like user profiles, review metadata, and contextual data can deepen the model's understanding and boost its predictive capabilities.

Moreover, for the model to be deployed in real-world situations, it needs to integrate smoothly with existing online platforms and review systems. The creation of intuitive interfaces and APIs for easy integration and scalability is vital for its widespread use and functionality.

Lastly, maintaining ongoing collaboration with industry stakeholders, regulatory bodies, and academic institutions is crucial to keep up with the emerging challenges and opportunities in the field of online review authenticity. By tackling these challenges and utilizing emerging technologies, the fraudulent review detection model can significantly contribute to promoting trust and integrity in online review ecosystems.

REFERENCES

[1] A. M. Elmogy, U. Tariq, A. Ibrahim, and A. Mohammed, "Fake Reviews Detection using Supervised Machine Learning," published in International Journal of Advanced Computer Science and Applications, 2021.

[2] W. H. Asaad, R. Allami, and Y. H. Ali, "Fake Review Detection Using Machine Learning,"by Revued'Intelligence Artificielle Vol. 37, No. 5, October 2023, pp. 1159-1166Journal ,25 July 2023.

[3] Mr. M. Thirunavukkarasu, P. Kavya, and M. P. T. Mahalakshmi, "Fake Reviews Detection using Supervised Machine Learning," April 2023.

[4] A. Moqueem, F. Moqueem, C. V. Reddy, D. Jayanth, and B. Brahma, "Online Shopping Fake Reviews Detection Using Machine Learning," Part of the book series: Communications in Computer and Information Science ((CCIS,volume 1697)), 01 January 2023.

[5] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jaraweh, and S. Maqsoon, "Fake Reviews Detection: A Survey," 26 April 2021.

[6] Ahsan, M.N.I., Nahian, T., Kafi, A.A., Hossain, M.d.I., Shah, F.M.: An ensemble approach to detect review spam using hybrid machine learning technique. In: 2016 19th

International Conference on Computer and Information Technology (ICCIT) (2016). https://doi.org/10.1109/iccitechn.2016.7860229

[7] Liu, W., He, J., Han, S., Cai, F., Yang, Z., Zhu, N.: A method for the detection of fake reviews based on temporal features of reviews and comments. IEEE Eng. Manag. Rev. 47(4), 67–79 (2019). https://doi.org/10.1109/EMR.2019.2928964

[8] Bali, A.P.S., Fernandes, M., Choubey, S., Goel, M.: Comparative performance of machine learning algorithms for fake news detection. In: Singh, M., Gupta, P.K., Tyagi, V., Flusser, J., Ören, T., Kashyap, R. (eds.) ICACDS 2019. CCIS, vol. 1046, pp. 420–430. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-9942-8_40

[9] M. Q. Bhat, D. S. Jayalakshmi, M. Mallegowda, and J. Geetha, "Interpreting Fake Reviews Using Machine Learning and Deep Learning," Part of the book series: Lecture Notes in Networks and Systems ((LNNS,volume 833)),01 March 2024.

[10] M. S. Jacob, S. Rajendran, V. M. Mario, K. T. Sai, and D. Logesh, "Fake Product Review Detection and Removal Using Opinion Mining Through Machine Learning," AISGSC 2019 2019: Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications pp 587–601, 13 March 2020.

[11] Seerat, B., Azam, F. (2012). Opinion mining: Issues and challenges (A survey). International Journal of Computer Applications, 49(9): 42-51.

[12] Rathor, A.S., Agarwal, A., Dimri, P. (2018) Comparative study of machine learning approaches for Amazon reviews. Procedia computer science, 132: 1552-1561. https://doi.org/10.1016/j.procs.2018.05.119

[13] Semchedine, M., Bensoula, N. (2022). Enhanced black widow algorithm for numerical functions optimization. Revue d'Intelligence Artificielle, 36(1): 1-11. https://doi.org/10.18280/ria.360101

[14] Al-Jarrah, M.A., Al-Jarrah, A., Jarrah, A., AlShurbaji, M., Magableh, S.K., Al-Tamimi, A.K., Bzoor, N., AlShamali, M.O. (2022). Accurate reader identification for the Arabic Holy Quran recitations based on an enhanced VQ algorithm. Revue d'Intelligence Artificielle, 36(6):815-823. https://doi.org/10.18280/ria.360601

[15] Nahma, D.R., Abbas, A.R. (2020). Patient opinion mining: Analysis of patient drugs satisfaction using support vector machine and logistic regression algorithm. Journal of Madenat Al-Elem College/Magallat Kulliyyat Madinat Al-ilm, 12(2): 164-171.

[16] Alameri SA, Mohd M (2021) Comparison of fake news detection using machine learning and deep learning techniques. In: 2021 3rd international cyber resilience conference (CRC), Langkawi Island, Malaysia, pp 1–6. https://doi.org/10.1109/CRC50527.2021.9392458