RESEARCH ARTICLE OPEN ACCESS

Improving Opinion Spam Identification with Sentiment Analysis using Explainable AI and Machine Learning

Apoorva Joshi[1], Namrata Jain[2]

[1,2]Computer Science & Engineering, Global Institute of Technology, Jaipur

ABSTRACT

Online reviews provide an essential effect on buying decisions, but the expansion of fake reviews and opinion spam challenges the trustworthiness of e-commerce. In addition to shifting customer perceptions misleading reviews damage competitor companies and erode confidence in marketplaces on the internet. They cause long-term instability in the field of e-commerce, yet they can instantly boost revenues for some. By integrating Explainable AI (XAI) with Sentiment Analysis (SA), this work enhances the detection of fake review. While XAI ensures transparency by providing an opinion on why a review is predicted as real or fake, SA helps in detecting suspicious emotional patterns in reviews. Consumer decisions are significantly affected by fake reviews, business reputation and necessitating effective detection mechanisms. We applied machine learning classifiers (Naïve Bayes, Random Forest, and Support Vector Machine (SVM)) and deep learning models (LSTM, BiLSTM, and CNN-LSTM) to analyze a dataset of 10,000 reviews. Our outcomes show that when it involves predicting untrue reviews, deep learning models perform better than ordinary classifiers. We used Local Interpretable Model-agnostic Explanations (LIME) to ensure that the detection process trustworthy and interpretable with the aim to maintain transparency. Also, we have predicted the product category with authenticity and sentiment of the review text using supervised learning techniques where Naïve Bayse performs well among SVM, Naïve Bayse & Random Forest. This method enhances fairness and reliability in online marketplaces while increasing the accuracy of fraud detection.

Keyword - Fake Review Detection, Opinion Spam, Sentiment Analysis, Explainable AI (XAI), Machine Learning, LSTM, Deep Learning, Opinion Manipulation, LIME, SHAP, Authenticity, Product Category Prediction

I. INTRODUCTION

The rapid growth of e-commerce has made online reviews a crucial factor influencing buyer decisions and brand reputations. However, the prevalence of fake reviews—either manually written or generated by AI-driven systems—threatens the credibility of online marketplaces. Deceptive reviews distort customer perceptions, unfairly impact business performance, and undermine trust in digital platforms**[1]**.

Sentiment analysis has emerged as a powerful tool for detecting fake reviews, as deceptive opinions often exhibit extreme polarities, with ratings skewed towards 1 (negative) or 5 (positive). These unnatural patterns suggest potential review manipulation, either to promote a product artificially or to harm a competitor's reputation**[2]**. While machine learning and deep learning models have significantly improved detection accuracy, creating challenges in understanding and justifying their decisions.

To address this, our research integrates sentiment analysis with Explainable AI (XAI) techniques, enhancing transparency in fake review detection**[3],[4]**. By applying Local Interpretable Model-agnostic Explanations (LIME), we provide insights into why a review is classified as fake or real, making the detection process more interpretable and trustworthy. Our study evaluates both traditional classification models—such as Naïve Bayes, Support Vector Machine (SVM), and Random Forest—and deep learning models—including Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and CNN-LSTM—to compare their effectiveness in identifying deceptive reviews.

In this research, we have trained a single model for both sentiment analysis and authenticity prediction of reviews using two different datasets. By utilizing these datasets, we aim to enhance the accuracy and reliability of the results. The paper is having various sections as Section II this section discusses existing research and highlights how certain companies generate fake reviews to manipulate consumer perceptions. Section III includes Methodology where, we outline the step-by-step process of developing our model, detailing the techniques and strategies used for sentiment and authenticity detection. Section IV This section provides an in-depth explanation of the models and classifiers used for training, including their implementation and configuration. In Section V We present the outcomes of our study, including evaluation metrics and a sample case study to demonstrate the model's effectiveness. Section VI Finally, we summarize our findings and discuss potential future improvements in fake review detection using advanced deep learning techniques.

By integrating sentiment analysis with fake review detection within a single model, our approach aims to provide a more comprehensive and accurate evaluation of online reviews.

II. LITERATURE REVIEW

A. *Fake Review Generation*

In recent years, a concerning trend has emerged where some trading companies and black-market asset groups are using social media platforms for deceptive practices. These entities create private chat groups or communities on platforms such as WhatsApp, Telegram, and Facebook, where they recruit individuals to perform specific tasks in exchange for financial incentives. The online reviews present on a website are expected to augment user credibility, attract consumer visits, and increase the hit ratio along with time spent on the site[5]. A major asset for users who are deciding to buy a product, watch a movie, or go to a restaurant and for managers who are making business decisions is Online reviews[6]. One of the most common tasks assigned to these members is the generation of fake reviews across various online platforms. These fraudulent schemes often target industries such as restaurants, e-commerce platforms, travel services, and local businesses. Participants are instructed to leave overly positive or negative reviews, irrespective of their actual experience with the product or service. If the reviews are divided into two rating categories — for example, 1-star and 5-star ratings — and among 500 reviews, 10 are 1-star while the remaining are 5-star, it indicates that most of the reviews are likely fake[7]. This manipulative practice misleads potential customers, skews product ratings, and creates an unfair advantage for businesses engaging in such deceitful tactics. Researchers in past have used different approaches to extract opinions. Broadly opinion mining can be extracted in two ways: supervised learning-based approach and deep learning-based approach.[8] This research aims to contribute to this fight against fake reviews by developing a single model capable of predicting both sentiment and authenticity, ensuring a more reliable and transparent online review system.

B. *Fake Review Detection: Deep Learning Models vs. Classification Models*

Traditional classification models, such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests, have been widely employed in detecting opinion spam[9]. These models typically rely on manually engineered features derived from review content, user behavior, or metadata. For instance, studies have utilized linguistic cues and reviewer activity patterns to identify suspicious reviews[10], [11].

However, the advent of deep learning has introduced models capable of automatically extracting complex patterns from data. Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Convolutional Neural Networks (CNNs) have demonstrated superior performance in capturing contextual nuances in textual data. A study proposed a deep learning-based framework that significantly improved detection accuracy by applying feature learning to behavioral features.

The Yelp Review Dataset was applied to supervised machine learning algorithms, In this study[12], with SVM obtaining a maximum accuracy of 0.899. In our research, we applied fake review detection on a labeled dataset for authenticity and sentiment analysis. The same fake review dataset is used for product category prediction also. SVM performed well in predicting sentiment as positive, negative, or neutral, while deep learning models performed well in detecting reviews as real or fake and Naïve Bayse performed well for product category prediction.

C. *Explainable AI (XAI)*

Explainable Artificial Intelligence (XAI) has emerged as a pivotal field of research aiming to create transparency in AI model predictions, thereby improving their interpretability. In the context of text classification, interpretability involves understanding why certain texts are classified into specific categories by the model. This understanding aids in uncovering the model's decision-making process and helps in trust building among its users[13].

While the deep learning models achieve high accuracy, interpretability concerns emerge driven by their black-box nature. In order to provide insights into model decisions, Explainable AI (XAI) includes techniques such as Local Interpretable Model-agnostic Explanations (LIME)[4], [14]. The usage of LIME in textual analysis improved the transparency of fake review detection in our research.

LIME works by generating perturbed versions of a given text sample, altering or removing certain words to observe changes in model predictions. It then trains a simpler, interpretable model (such as a linear classifier) on these perturbed instances to approximate how the machine learning model makes decisions. By analyzing the impact of individual words on predictions, LIME provides a visual representation of the most impactful terms, helping users understand why a review was classified as real or fake.

LIME is not used for topic modelling directly in our study, but it rather provides complementary local explanations that, combined with the semantic topics from Top2Vec, give a richer understanding of the [15]model's fake review detection logic. Specifically, LIME is used to analyze the importance of local features for the model's fake review predictions.

As the author have used SHAP and Counterfactual Explanation in Paper[9]. We applied LIME in our research to detect review authenticity and classification of sentiment. Two classes are used to train the model: ["Real", "Fake"] for review classification and ["Negative", "Positive"] for sentiment analysis. LIME helped identify key words driving sentiment predictions while also highlighting suspicious terms in fake reviews, enhancing interpretability and trust in our detection system[16].

III. METHODOLOGY

As Figure 1 proposed our methodology for fake review detection using sentiment analysis begins with dataset selection, where two different labeled datasets are utilized to train separate models. The Yelp Review Dataset is used for sentiment prediction[17] and the fake review dataset is used for authentic prediction of an opinion. The datasets undergo preprocessing, which includes cleaning, normalization, vectorization and tokenization of text data to remove noise and standardize inputs. Feature extraction is then performed to derive meaningful representations from the text data, such as word embeddings, sentiment scores, and metadata features.

Two parallel processes are carried out: one focusing on deep learning for sentiment analysis and review prediction as real or fake. Another includes supervised classification algorithms to train the model for review prediction in terms of sentiment and fake or real. For classification, six different models—LSTM, BiLSTM, CNN-LSTM, Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB)—are employed for both sentiment analysis and authenticity detection[18]. The deep learning models (LSTM, BiLSTM, CNN-LSTM) leverage sequential dependencies in textual data, while the traditional machine learning models (SVM, RF, NB) provide complementary approaches based on statistical and probabilistic techniques. Determine whether a review is positive/negative/neutral and real/fake.

Standard evaluation metrics such as accuracy, precision, recall, and F1-score are used to evaluate each model's performance[10]. A comparative analysis is conducted to assess the effectiveness of different models across both classification tasks. Explainable AI (XAI) techniques, specifically Local Interpretable Model-agnostic Explanations (LIME), are applied to highlight influential features and explain model decisions to enhance interpretability. The final results are analyzed to derive insights, concluding the methodology.

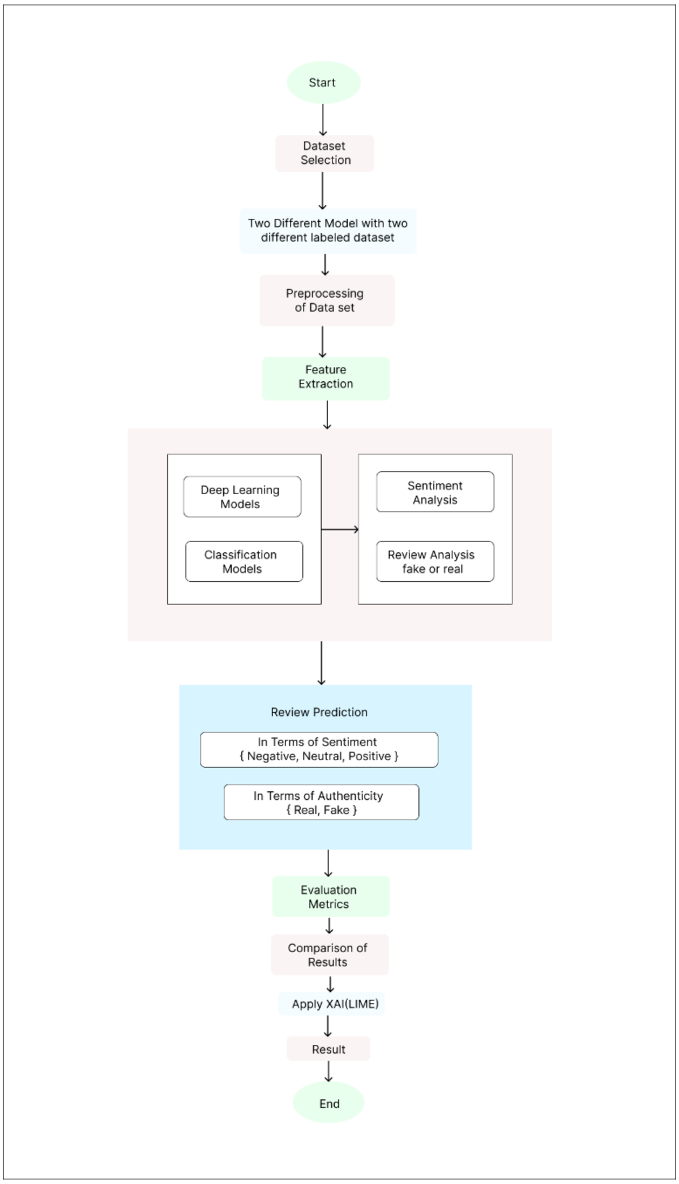

Start
Dataset Selection
Two Different Model with two different labeled dataset
Preprocessing of Data set
Feature Extraction
Deep Learning Models
Classification Models
Sentiment Analysis
Review Analysis fake or real
Review Prediction
In Terms of Sentiment { Negative, Neutral, Positive }
In Terms of Authenticity { Real, Fake }
Evaluation Metrics
Comparison of Results
Apply XAI(LIME)
Result
End

Figure 1 – Methodology

IV. EXPERIMENTAL SETUP

A. ***Dataset Selection*** - Our research is conducted on two distinct labeled datasets: 1) Yelp Review Dataset for sentiment analysis and 2) Fake Review Dataset for fake or real review prediction. Both are respectively labeled as sentiment {0 =Negative, 1 = Positive}, authenticity {0 = Fake, 1 = Real}.

B. ***Data Preprocessing*** - The data preprocessing process begins with text cleaning, which is an essential step to standardize the textual data before feature extraction and model training. The cleaning process follows these key steps: Lowercasing, Removing Numbers (This prevents numerical values from affecting the text-based learning process). Punctuation Removal, Stripping Whitespace[19]. This preprocessing step is crucial in removing noise from the data and improving the performance of both deep learning and machine learning models used for sentiment classification and authenticity detection.

C. ***Feature Engineering*** **-** Labels are converted to integers for model compatibility, and TF-IDF (Term Frequency-Inverse Document Frequency) is applied to extract meaningful features from text, limiting vocabulary to 5000 words and removing stop words to enhance model performance[13], [20].

D. ***Dataset Splitting*** **-** The datasets for sentiment analysis and fake review detection are split into training (80%) and testing (20%) sets using "train_test_split", ensuring consistent randomization with {random_state=42} for reproducibility.

E. ***Machine Learning Algorithms for Model Training*** **-** For sentiment analysis and fake review detection, both supervised machine learning techniques and deep learning models are employed. The supervised learning techniques include Naïve Bayes, Random Forest, and Support Vector Machine (SVM)**,** which leverage statistical and rule-based approaches to classify text. Among them, as shown in Table-1 SVM achieved the highest accuracy of 0.89 for sentiment analysis and 0.85 for fake review detection, demonstrating its effectiveness in handling text classification tasks. Naïve Bayes and Random Forest also performed well, with accuracies around 0.86 for sentiment analysis and slightly lower for fake review detection.

Deep learning models, including LSTM, BiLSTM, and CNN-LSTM, utilize neural networks to learn complex patterns in textual data. These models process text as sequential input, capturing contextual dependencies that traditional machine learning methods might miss. As shown in, Table -1 BiLSTM outperformed all models in fake review detection with an accuracy of 0.9337, followed by LSTM (0.9243) and CNN-LSTM (0.9147). In sentiment analysis, CNN-LSTM achieved the highest accuracy (0.8797), followed closely by LSTM (0.8753) and BiLSTM (0.8443). These results highlight that deep learning models, particularly BiLSTM, excel in detecting fake reviews, while SVM remains highly competitive in sentiment classification. Naive Bayes achieved the highest overall accuracy (75%), but SVM provided more balanced results across multiple categories, making it a strong candidate for practical use.

F. ***Applying XAI(LIME)*** **-** For LIME explanation a function first tokenizes the input text using Keras' Tokenizer, converts it into sequences, and applies padding to match the required input shape. A wrapper function processes the text and returns model predictions, ensuring that for binary classification, single-node outputs are converted into a two-class probability format. LIME then generates explanations by perturbing the input and analyzing the model's behavior, presenting the most influential words in a notebook visualization. This approach helps in understanding how deep learning models, powered by neural networks, make decisions, thereby improving transparency and trust in their predictions[4], [21].

As mentioned in Figure 3 when LIME is applied to explain that why the review is predicted as Negative/Positive and Fake/Real the LSTM model performs highly accurate with the accuracy for 0.99 in case of sentiment and 0.97 in case or authenticity, the impactful words for the review text "The product was amazing! I absolutely loved it." was {amazing, product, loved, absolutely}.

V. RESULT

We applied supervised learning algorithms and deep learning algorithms on two different datasets for review prediction. One dataset was used for sentiment analysis, while the other was used for authenticity detection. The first dataset was labeled with polarity {0,1}, and the second dataset was labeled as CG (Computer Generated) or OG (Original)[3], [8]. We replaced CG and OG with their numeric representations, 0 and 1, respectively. After applying the machine learning models, the results were observed as mentioned in this section. Table 1 presents the accuracy of different models for both Sentiment Analysis and Fake Review Detection. The following observations can be made:

A. ***Sentiment Analysis***-The SVM model achieved the highest accuracy at 0.89, followed by CNN-LSTM (0.8797) and LSTM (0.8753). Traditional machine learning models like Naïve Bayes (0.86) and Random Forest (0.86) performed similarly but were outperformed by deep learning models. BiLSTM (0.8443) had the lowest accuracy among deep learning models, indicating that bidirectional learning did not significantly improve sentiment prediction in this case.

B. ***Fake Review Detection***-The BiLSTM model achieved the highest accuracy (0.9337), followed closely by LSTM (0.9243) and CNN-LSTM (0.9147). Among machine learning models, SVM (0.85) performed the best, whereas Random Forest (0.81) had the lowest accuracy. The deep learning models consistently outperformed traditional models in Fake Review Detection, indicating the benefit of sequential learning for authenticity assessment.

These findings suggest that SVM is the best choice for sentiment analysis, while BiLSTM is the most effective for detecting fake reviews.

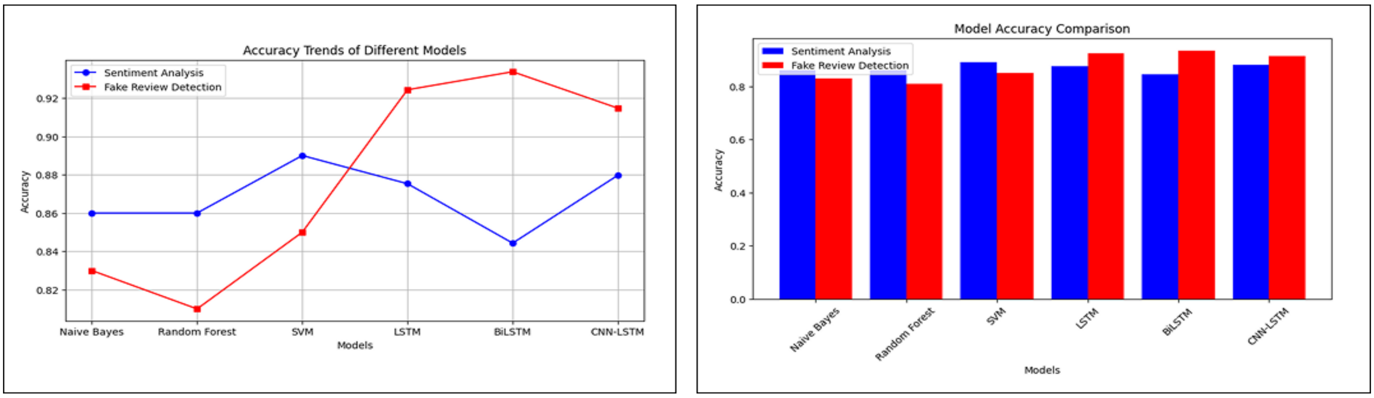
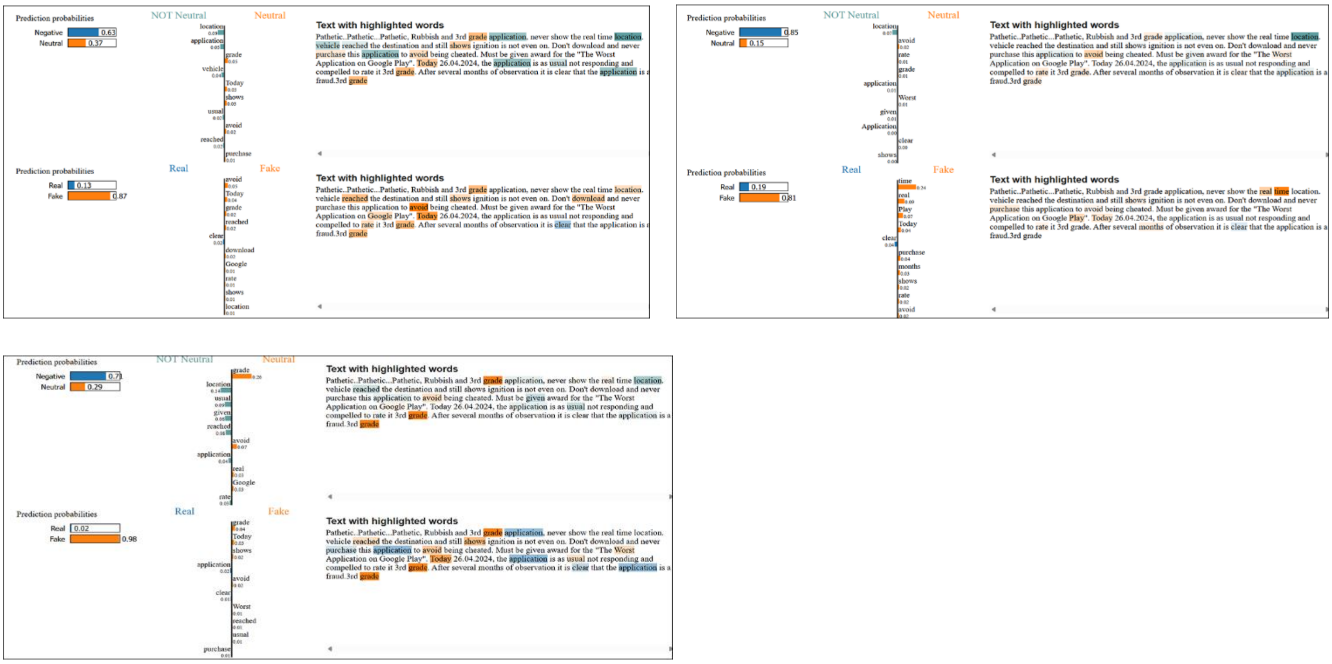


Figure 2- Distinct Model Accuracy using Line Chart and Bar Chart



Text with highlighted words

Pathetic..Pathetic...Pathetic, Rubbish and 3rd grade application, never show the real time location. vehicle reached the destination and still shows ignition is not even on. Don't download and never purchase this application to avoid being cheated. Must be given award for the "The Worst Application on Google Play". Today 26.04.2024, the application is as usual not responding and compelled to rate it 3rd grade. After several months of observation it is clear that the application is a fraud.3rd grade

Text with highlighted words

Pathetic..Pathetic...Pathetic, Rubbish and 3rd grade application, never show the real time location. vehicle reached the destination and still shows ignition is not even on. Don't download and never purchase this application to avoid being cheated. Must be given award for the "The Worst Application on Google Play". Today 26.04.2024, the application is as usual not responding and compelled to rate it 3rd grade. After several months of observation it is clear that the application is a fraud.3rd grade

Text with highlighted words

Pathetic..Pathetic...Pathetic, Rubbish and 3rd grade application, never show the real time location. vehicle reached the destination and still shows ignition is not even on. Don't download and never purchase this application to avoid being cheated. Must be given award for the "The Worst Application on Google Play". Today 26.04.2024, the application is as usual not responding and compelled to rate it 3rd grade. After several months of observation it is clear that the application is a fraud.3rd grade

Text with highlighted words

Pathetic..Pathetic...Pathetic, Rubbish and 3rd grade application, never show the real time location. vehicle reached the destination and still shows ignition is not even on. Don't download and never purchase this application to avoid being cheated. Must be given award for the "The Worst Application on Google Play". Today 26.04.2024, the application is as usual not responding and compelled to rate it 3rd grade. After several months of observation it is clear that the application is a fraud.3rd grade

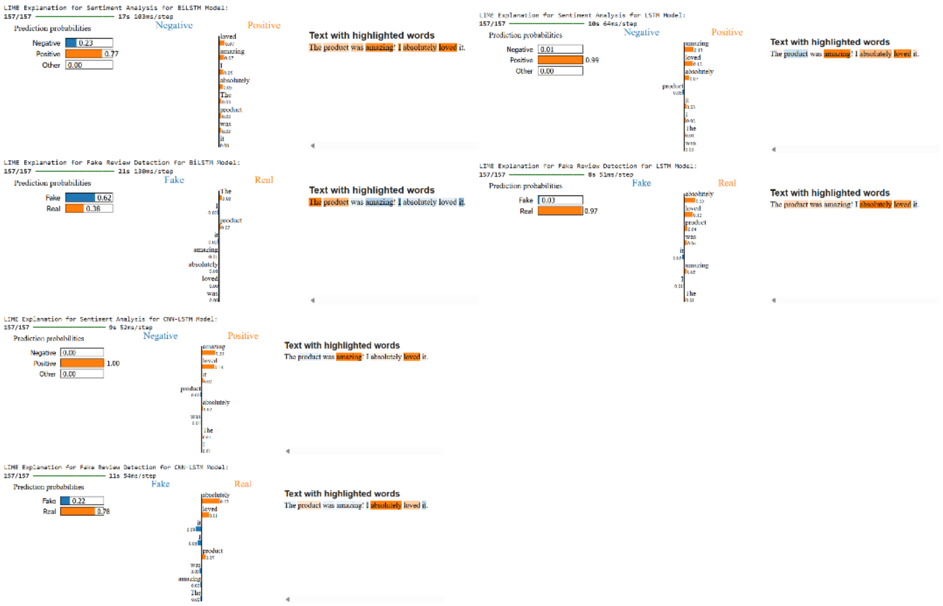
Text with highlighted words

Pathetic..Pathetic...Pathetic, Rubbish and 3rd grade application, never show the real time location. vehicle reached the destination and still shows ignition is not even on. Don't download and never purchase this application to avoid being cheated. Must be given award for the "The Worst Application on Google Play". Today 26.04.2024, the application is as usual not responding and compelled to rate it 3rd grade. After several months of observation it is clear that the application is a fraud.3rd grade

Text with highlighted words

Pathetic..Pathetic...Pathetic, Rubbish and 3rd grade application, never show the real time location. vehicle reached the destination and still shows ignition is not even on. Don't download and never purchase this application to avoid being cheated. Must be given award for the "The Worst Application on Google Play". Today 26.04.2024, the application is as usual not responding and compelled to rate it 3rd grade. After several months of observation it is clear that the application is a fraud.3rd grade

Figure 3- LIME Implementation for Supervised Learning Algorithms ("Naïve Bayes", "Random Forest", "SVM")



```
LIME Explanation for Sentiment Analysis for BiLSTM Model:
157/157 ━━━━━━━━━━━━━━━━━━━━ 17s 103ms/step
```

Text with highlighted words

The product was amazing! I absolutely loved it.

```
LIME Explanation for Fake Review Detection for BiLSTM Model:
157/157 ━━━━━━━━━━━━━━━━━━━━ 21s 130ms/step
```

Text with highlighted words

The product was amazing! I absolutely loved it.

```
LIME Explanation for Sentiment Analysis for CNN-LSTM Model:
157/157 ━━━━━━━━━━━━━━━━━━━━ 9s 52ms/step
```

Text with highlighted words

The product was amazing! I absolutely loved it.

```
LIME Explanation for Fake Review Detection for CNN-LSTM Model:
157/157 ━━━━━━━━━━━━━━━━━━━━ 11s 54ms/step
```

Text with highlighted words

The product was amazing! I absolutely loved it.

```
LIME Explanation for Sentiment Analysis for LSTM Model:
157/157 ━━━━━━━━━━━━━━━━━━━━ 10s 64ms/step
```

Text with highlighted words

The product was amazing! I absolutely loved it.

```
LIME Explanation for Fake Review Detection for LSTM Model:
157/157 ━━━━━━━━━━━━━━━━━━━━ 8s 51ms/step
```

Text with highlighted words

The product was amazing! I absolutely loved it.

Figure 4- LIME Implementation for Deep Learning Algorithms (“LSTM”, “BiLSTM”, “CNN-LSTM”)

Table 1 - DISTINCT MODEL ACCURACY FOR SENTIMENT AND AUTHENTICITY PREDICTION

Index	Model	Sentiment Analysis Accuracy	Fake Review Detection Accuracy
1	Naive Bayes	0.86	0.83
2	Random Forest	0.86	0.81
3	SVM	0.89	0.85
4	LSTM	0.8753	0.9243
5	BiLSTM	0.8443	0.9337
6	CNN-LSTM	0.8797	0.9147

Table 2 - SENTIMENT ANALYSIS EVALUATION METRICS (NAÏVE BAYES)

Class	Precision	Recall	F1- Score	Support
Negative (0)	0.87	0.86	0.87	1511
Positive (1)	0.86	0.87	0.86	1489
Accuracy	-	-	0.86	3000
Macro Avg	0.87	0.87	0.86	3000
Weighted Avg	0.87	0.86	0.87	3000

Table 3 - SENTIMENT ANALYSIS EVALUATION METRICS (RANDOM FOREST)

Class	Precision	Recall	F1- Score	Support
Negative (0)	0.85	0.87	0.86	1511
Positive (1)	0.87	0.84	0.85	1489
Accuracy	-	-	0.86	3000
Macro Avg	0.86	0.86	0.86	3000
Weighted Avg	0.86	0.86	0.86	3000

Table 4 - SENTIMENT ANALYSIS EVALUATION METRICS (SVM)

Class	Precision	Recall	F1- Score	Support
Negative (0)	0.90	0.89	0.89	1511
Positive (1)	0.89	0.90	0.89	1489
Accuracy	-	-	0.89	3000
Macro Avg	0.89	0.89	0.89	3000
Weighted Avg	0.89	0.89	0.89	3000

Table 5 - FAKE REVIEW DETECTION EVALUATION METRICS (NAÏVE BAYES)

Class	Precision	Recall	F1- Score	Support
Fake (0)	0.80	0.88	0.84	1527
Real (1)	0.87	0.78	0.82	1473
Accuracy	-	-	0.83	3000
Macro Avg	0.83	0.83	0.83	3000
Weighted Avg	0.83	0.83	0.83	3000

Table 6 - FAKE REVIEW DETECTION EVALUATION METRICS (RANDOM FOREST)

Class	Precision	Reca ll	F1- Score	Support
Fake (0)	0.79	0.87	0.83	1527
Real (1)	0.85	0.76	0.80	1473
Accuracy	-	-	0.81	3000
Macro Avg	0.82	0.81	0.81	3000

Weighted Avg	0.82	0.81	0.81	3000

Table 7 - FAKE REVIEW DETECTION EVALUATION METRICS (SVM)

Class	Precision	Recall	F1- Score	Support
Fake (0)	0.86	0.85	0.86	1527
Real (1)	0.85	0.86	0.85	1473
Accuracy	-	-	0.85	3000
Macro Avg	0.85	0.85	0.85	3000
Weighted Avg	0.85	0.85	0.85	3000

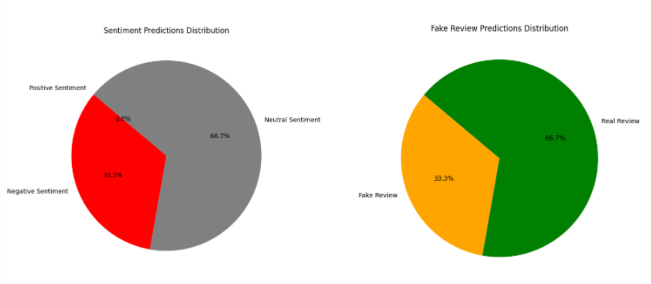
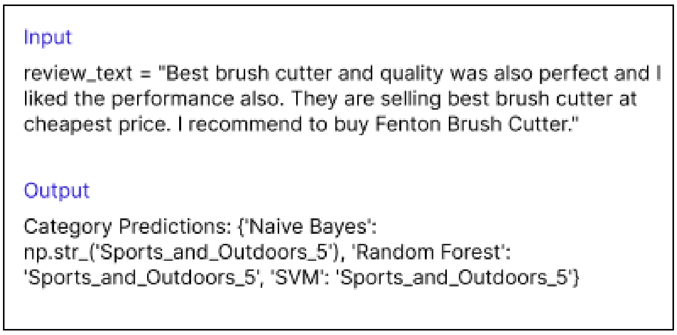


Figure 5- Sentiment and Authenticity Prediction for a Particular review



Input

review_text = "Best brush cutter and quality was also perfect and I liked the performance also. They are selling best brush cutter at cheapest price. I recommend to buy Fenton Brush Cutter."

Output

Category Predictions: {'Naive Bayes': np.str_('Sports_and_Outdoors_5'), 'Random Forest': 'Sports_and_Outdoors_5', 'SVM': 'Sports_and_Outdoors_5'}

Figure 6- Prediction of Product Category for a Particular revie

Table 8 - Classification Report of Category Prediction using Naïve Bays

Class	Precision	Recall	F1-Score	Support
Books_5	0.74	0.56	0.64	327
Clothing_Shoes_and_Jewelry_5	0.76	0.81	0.78	300
Electronics_5	0.78	0.79	0.79	306
Home_and_Kitchen_5	0.66	0.70	0.68	316
Kindle_Store_5	0.64	0.88	0.74	324
Movies_and_TV_5	0.93	0.83	0.88	271
Pet_Supplies_5	0.83	0.84	0.84	308
Sports_and_Outdoors_5	0.66	0.57	0.61	296
Tools_and_Home_Improvement_5	0.70	0.67	0.69	267
Toys_and_Games_5	0.85	0.81	0.83	285
Accuracy	-	-	0.75	3000
Macro avg	0.75	0.75	0.75	3000
Weighted avg	0.75	0.75	0.75	3000

Table 9 - Classification Report of Category Prediction using Random Forest

Class	Precision	Recall	F1-Score	Support
Books_5	0.70	0.65	0.67	327
Clothing_Shoes_and_Jewelry_5	0.61	0.75	0.67	300
Electronics_5	0.70	0.67	0.68	306
Home_and_Kitchen_5	0.54	0.55	0.55	316
Kindle_Store_5	0.69	0.73	0.71	324
Movies_and_TV_5	0.79	0.86	0.83	271
Pet_Supplies_5	0.86	0.77	0.81	308
Sports_and_Outdoors_5	0.50	0.40	0.44	296
Tools_and_Home_Improvement_5	0.57	0.58	0.57	267
Toys_and_Games_5	0.76	0.76	0.76	285
Accuracy	-	-	0.67	3000
Macro avg	0.67	0.67	0.67	3000
Weighted avg	0.67	0.67	0.67	3000

Table 10 - Classification Report of Category Prediction using SVM

Class	Precision	Recall	F1-Score	Support
Books_5	0.74	0.69	0.72	327
Clothing_Shoes_and_Jewelry_5	0.78	0.76	0.77	300
Electronics_5	0.74	0.74	0.74	306
Home_and_Kitchen_5	0.61	0.70	0.65	316
Kindle_Store_5	0.73	0.77	0.75	324
Movies_and_TV_5	0.90	0.85	0.88	271
Pet_Supplies_5	0.90	0.80	0.84	308
Sports_and_Outdoors_5	0.54	0.60	0.57	296
Tools_and_Home_Improvement_5	0.67	0.69	0.68	267
Toys_and_Games_5	0.83	0.79	0.81	285
Accuracy	-	-	0.74	3000
Macro avg	0.75	0.74	0.74	3000
Weighted avg	0.74	0.74	0.74	3000

Figure 2 presents the Sentiment Predictions Distribution and Fake Review Predictions Distribution using pie charts for a particular new review. Sentiment Analysis: Neutral sentiments dominate (66.7%), followed by negative sentiments (33.3%), while positive sentiment is absent (0%) in the given dataset visualization. Fake Review Detection: 66.7% of reviews are classified as real, while 33.3% are detected as fake.

Figure 3 visualizes the accuracy trends of different models for fake review detection using sentiment analysis. SVM outperforms all other models, achieving the highest accuracy (0.89). LSTM and CNN-LSTM also show strong performance, with both achieving accuracies above 0.87. Supervised machine learning models like Naïve Bayes and Random Forest remain consistent at 0.86, indicating their reliability but also their limitations compared to deep learning approaches. The performance of deep learning models is significantly higher, with BiLSTM reaching the peak accuracy of 0.9337, followed by LSTM (0.9243) and CNN-LSTM (0.9147). Supervised learning models struggle to reach similar performance, with SVM being the best among them at 0.85.

The sharp increase in accuracy for deep learning models (especially from LSTM onwards) highlights the importance of neural networks in improving fake review detection.

Table 2-7 shows the evaluation metrics for ex - accuracy, precision, recall, f1-score for all the supervised learning algorithms which are used to train our model. The performance evaluation of sentiment analysis and fake review detection models demonstrates varying strengths across different algorithms. For sentiment analysis, Naïve Bayes achieves an accuracy of 86% with balanced performance across both negative and positive sentiment classes, reflected by its F1-scores of 0.87 and 0.86, respectively as shown in Table 2. Similarly, Table 3 demonstrates that Random Forest matches this 86% accuracy with comparable precision and recall values, indicating consistent performance. However, Table 4 shows SVM stands out with the highest accuracy of 89%, maintaining an F1-score of 0.89 for both classes, showcasing its superior ability to classify sentiment with remarkable consistency.

The performance of three models — Naive Bayes, Random Forest, and SVM — was evaluated for predicting product categories using key metrics such as precision, recall, f1-score, and accuracy. Among the models, Naive Bayes achieved the highest overall accuracy of 75%, with its best performance in the Movies_and_TV_5 category (f1-score of 0.88) as shown in Table-8. However, it struggled with Sports_and_Outdoors_5 (f1-score of 0.61), indicating challenges in certain categories. Random Forest, with an overall accuracy of 67%, performed best in Movies_and_TV_5 (f1-score of 0.83) but exhibited lower performance in Sports_and_Outdoors_5 (f1-score of 0.44), suggesting underfitting in complex or less frequent categories as demonstrated in Table-9. Table 10 shows the evaluation matrices for SVM such as SVM achieved an overall accuracy of 74%, comparable to Naive Bayes, with strong results in Movies_and_TV_5 (f1-score of 0.88) and Pet_Supplies_5 (f1-score of 0.84). However, it faced moderate issues with Sports_and_Outdoors_5 (f1-score of 0.57). Overall, Naive Bayes and SVM outperformed Random Forest, particularly in handling a broader range of categories. Movies_and_TV_5 consistently had the highest scores across all models, indicating distinct textual features, while Sports_and_Outdoors_5 remained the most challenging category due to possible overlap in textual characteristics. SVM demonstrated balanced performance across multiple categories, making it a competitive choice for practical applications.

In the context of fake review detection, Naïve Bayes achieves an accuracy of 83%, with a slightly better recall for the fake class (0.88) compared to the real class (0.78), resulting in a slight performance bias towards identifying fake reviews as shown in Table 5. In Table 6, Random Forest, on the other hand, records a slightly lower accuracy of 81% with an F1-score of 0.83 for fake reviews and 0.80 for real reviews, reflecting a reasonable yet modest performance. SVM once again emerges as the top performer with an accuracy of 85% and balanced F1-scores of 0.86 and 0.85 for fake and real reviews, respectively as mentioned in Table 7. This indicates SVM's superior capability in handling the nuanced task of distinguishing between genuine and fake reviews.

Overall, SVM consistently outperforms the other two models in both sentiment analysis and fake review detection, offering the highest accuracy and balanced precision, recall, and F1-scores. Naïve Bayes delivers competitive results, particularly in sentiment analysis, while Random Forest maintains steady performance without surpassing the other models. These findings highlight the effectiveness of SVM for both sentiment analysis and authenticity detection tasks.

In Figure 6, the input demonstrates a sample review text describing a product — specifically a "brush cutter" — with positive feedback on its quality and performance. The review emphasizes user satisfaction and includes a purchase recommendation. The output section presents the predictions generated by multiple machine learning models (Naive Bayes, Random Forest, and SVM) trained for category classification. Each model correctly identifies the category of the product as "Sports_and_Outdoors_5", indicating consistency across models. The use of TF-IDF vectorization for feature extraction likely helped capture relevant textual patterns related to the product's category. This showcases the models' ability to leverage textual data effectively for accurate classification.

In Figure 4, LIME explanation is determined by highlighting the most impactful word. The LSTM performs well as the review prediction accuracy of it is highest for fake or real classes and CNN-LSTM performs highly accurate in polarity prediction is 1.

VI. CONCLUSION

Among traditional models, SVM demonstrates the best performance for sentiment analysis, while BiLSTM proves to be the most effective for detecting fake reviews. Deep learning models outperform traditional machine learning models in fake review detection, showcasing their ability to capture complex textual patterns. Overall, SVM consistently outperforms the other two models in both sentiment analysis and fake review detection, offering the highest accuracy and balanced precision, recall, and F1-scores. Naïve Bayes delivers competitive results, particularly in sentiment analysis, while Random Forest maintains steady performance without surpassing the other models. These findings highlight the effectiveness of SVM for both sentiment analysis and authenticity detection tasks. The high percentage of neutral sentiment in the dataset indicates the need for a more balanced dataset to improve sentiment classification. Naive Bayes achieved the highest overall accuracy (75%), but SVM provided more balanced results across multiple categories, making it a strong candidate for practical use. Further tuning, such as hyperparameter optimization or ensemble models, could improve performance, especially for underperforming categories. These findings validate the effectiveness of deep learning architectures in text-based classification tasks and highlight the importance of neural networks for fake review detection.

Additionally, since deep learning architectures leverage neural networks for textual analysis, LIME-based explain ability methods yield more efficient and meaningful results. Finally, we demonstrate that although explain-ability features associated with LIME suffer from poor reputation due to restricted interpretation capability, if applied strategically, it can be used to measure and optimize LSTM model performance against LSTM model explain ability. For future work, more deep learning techniques such as BERT and RoBERTa can be explored for model training. In the context of Explainable AI (XAI), counterfactual explanations and SHAP can be emphasized to enhance interpretability.

ACKNOWLEDGEMENT

The authors express their gratitude for the invaluable support received during this research. I am especially grateful to my guide for their guidance throughout the process of publishing this research paper. I also extend my appreciation to my institute for organizing the conference and providing us with the opportunity to present and publish our work.

REFERENCES

[1] M. Shajalal, M. Atabuzzaman, A. Boden, G. Stevens, and D. Du, “What Matters in Explanations: Towards Explainable Fake Review Detection Focusing on Transformers,” Jul. 24, 2024, *arXiv*: arXiv:2407.21056. doi: 10.48550/arXiv.2407.21056.

[2] Z. Singla, S. Randhawa, and S. Jain, “Statistical and sentiment analysis of consumer product reviews,” in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi: IEEE, Jul. 2017, pp. 1–6. doi: 10.1109/ICCCNT.2017.8203960.

[3] O. Chernyaeva, T. Hong, and O.-K. D. Lee, “Deconstructing Review Deception: A Study on Counterfactual Explanation and

XAI in Detecting Fake and GPT-Generated Reviews," presented at the Hawaii International Conference on System Sciences, 2024. doi: 10.24251/HICSS.2024.056.

[4] R. Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," ACM Comput. Surv., vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.

[5] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," Inf. Sci., vol. 311, pp. 18–38, Aug. 2015, doi: 10.1016/j.ins.2015.03.040.

[6] P. Anil, R. J. V. Siddardha, and T. Antony, "FAKE REVIEW SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING," vol. 12, no. 7.

[7] Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[8] F. Cardoso, R. M. Silva, and T. A. Almeida, "Towards automatic filtering of fake reviews," Neurocomputing, vol. 309, pp. 106–116, Oct. 2018, doi: 10.1016/j.neucom.2018.04.074.

[9] Y. Wu, E. W. T. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," Decis. Support Syst., vol. 132, p. 113280, May 2020, doi: 10.1016/j.dss.2020.113280.

[10] Ejaz, Z. Turabee, M. Rahim, and S. Khoja, "Opinion mining approaches on Amazon product reviews: A comparative study," in 2017 International Conference on Information and Communication Technologies (ICICT), Karachi: IEEE, Dec. 2017, pp. 173–179. doi: 10.1109/ICICT.2017.8320185.

[11] N. A. Patel and R. Patel, "A Survey on Fake Review Detection using Machine Learning Techniques," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India: IEEE, Dec. 2018, pp. 1–6. doi: 10.1109/CCAA.2018.8777594.

[12] R. Das et al., "Towards the development of an explainable e-commerce fake review index: An attribute analytics approach," Eur. J. Oper. Res., vol. 317, no. 2, pp. 382–400, Sep. 2024, doi: 10.1016/j.ejor.2024.03.008.

[13] J. Prager, "Open-Domain Question–Answering," Found. Trends® Inf. Retr., vol. 1, no. 2, pp. 91–231, 2006, doi: 10.1561/1500000001.

[14] Cambria, Erik, Björn Schuller, Yunqing Xia, and Catherine Havasi. "New avenues in opinion mining and sentiment analysis." IEEE Intelligent systems 28, no. 2 (2013): 15-21.

[15] K. Sharma, R. Ajmera, and D. K. Dharamdasani, "Effect of number of processor on the cache hit rate in symmetric multiprocessor environment," Journal of Discrete Mathematical Sciences and Cryptography, vol. 22, no. 4, pp. 509–520, May 2019,

[16] S. A. Saiyed, N. Sharma, H. Kaushik, P. Jain, G. K. Soni and R. Joshi, "Transforming portfolio management with AI and ML: shaping investor perceptions and the future of the Indian investment sector," Parul University International Conference on Engineering and Technology 2025 (PiCET 2025), pp. 1108-1114, 2025.

[17] Gautam, A, R. Ajmera, D. K. Dharamdasani, S. Srivastava, and A. Johari, "Improving climate change predictions using time series analysis and deep learning," Global and Stochastic Analysis, vol. 12, no. 4, Jul. 2025.

[18] Kaushik, H. Arora, R. Joshi, K. Sharma, M. Mehra and P. K. Sharma, "Digital Image Security using Hybrid Model of Steganography and Cryptography," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 1009-1012

[19] N. Soni, N. Nigam, "Recent Advances in Artificial Intelligence and Machine Learning: Trends, Challenges, and Future

Directions", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 1, pp. 9-12, 2025.

[20] Kaushik, "Artificial Intelligence: Recent Advances, Challenges, and Future Directions", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 2, 2025.

[21] Maheshwari, R. Ajmera, and D. K. Dharamdasani, "Unmasking Embedded Text: A Deep Dive into Scene Image Analysis," in 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India: IEEE, Nov. 2023, pp. 1403–1408.

[22] P. Jha, T. Biswas, U. Sagar and K. Ahuja, "Prediction with ML paradigm in Healthcare System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1334-1342, 2021.

[23] Jha, P., Dembla, D. & Dubey, W. Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model. Multimed Tools Appl 83, 37839–37858 (2024).

[24] P. Jha, D. Dembla and W. Dubey, "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 569-574, 2023.

[25] Jha, P., Dembla, D., Dubey, W., "Crop Disease Detection and Classification Using Deep Learning-Based Classifier Algorithm", Emerging Trends in Expert Applications and Security. ICETEAS 2023. Lecture Notes in Networks and Systems, vol 682. 2023.

[26] P. Jha, M. Mathur, A. Purohit, A. Joshi, A. Johari and S. Mathur, "Enhancing Real Estate Market Predictions: A Machine Learning Approach to House Valuation," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1930-1934, 2025.

[27] Gaur, P., Vashistha, S., Jha, P. (2023). "Twitter Sentiment Analysis Using Naive Bayes-Based Machine Learning Technique", Sentiment Analysis and Deep Learning. Advances in Intelligent Systems and Computing, vol 1432.

[28] P. Upadhyay, K. K. Sharma, R. Dwivedi and P. Jha, "A Statistical Machine Learning Approach to Optimize Workload in Cloud Data Centre," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 276-280.