Fake Review Detection System

Hrudhvik Nangineni
B.Tech CSE
SCOPE
VIT Chennai

S Geetha
Professor and Associate Dean
SCOPE
VIT Chennai

***Abstract*—In this era of internet, e-commerce has grown tremendously. The customers are increasingly relying on reviews for product information. However, the usefulness of these online reviews is hindered by fake reviews which give misleading information about the product. These reviews are written in an agenda to increase the product sale by misleading the customer. As these reviews influence the purchasing ability of the future customer, it can give a positive or negative impact on the businesses. Fake reviews can not only impact the reputation of the businesses but also involve financial losses as well. Therefore, detection of fake reviews is needed to solve the problem for maintaining the integrity of online reviews. To overcome the traditional machine learning model, here we propose to find a fake review detection by applying deep neural networks Bi-LSTM, DistilBERT, Xlnet, ALBERT, DeBERTa and finetuned distilBERT models and perform a comparative analysis on the performance of the models. The proposed models gave a promising performance for detecting the fake reviews. The transformer-based models like DistilBERT, AlBERT, and XlNET displayed high accuracy, precision, recall, and F1-score were repeatedly, indicating their potential for usage in real-world situations involving the detection of false reviews in many domains.**

***Index Terms*—neural network, deep learning, fake review detection.**

I. Introduction

In the age of e-commerce and online buying, the problem of fake reviews has spread widely. Fake reviews can have a significant impact on customer behaviour and ultimately damage the credibility of the platform because consumers are relying more and more on online platforms to make purchasing decisions. A fake review is one that has been created or altered artificially with the intention of deceiving customers. This can be done for a number of reasons, such as increasing sales, damaging the reputation of a rival, or just for one's own benefit.

Artificial intelligence (AI)-based fake review detection systems have been presented as a solution to this problem. Based on several characteristics, including writing style, sentiment, and consistency with other reviews, these systems utilise machine learning algorithms to assess reviews and identify false reviews. Artificial intelligence (AI) solutions have the potential to dramatically increase the accuracy and efficiency of fake review detection compared to manual techniques by automating the process. Natural language processing is one of the most widely employed methods by AI-based fake review identification systems . In order to evaluate whether a review is real or fraudulent, NLP algorithms are taught to examine the language used, including the words used, sentence structure, and sentiment. Graph analysis, which employs network analysis methods to find patterns of behaviour in the relationships between reviews, reviewers, and products, and sentiment analysis, which employs machine learning algorithms to analyse the sentiment expressed in a review, are additional techniques that can be used. Although AI-based fake review detection systems have some promise, there are still a number of issues that need to be resolved. The necessity for a lot of high-quality training data to train the algorithms is one of the key difficulties. Systems that can deal with the dynamic nature of fraudulent reviews, which might change continuously in order to avoid detection, present another issue. In our study, we examine the effectiveness of multiple deep learning models across diverse dataset types and sizes. The study aims to clarify the performance variations between these models and offer perceptions about their applicability for diverse sorts of data.

A. Objectives

- To experiment with deep neural network model Bidirectional-LSTM for detecting fake reviews.
- To experiment with Transformations models DistilBERT, Xlnet, ALBERT, DeBERTa and finetuned DistilBERT for detecting fake reviews.
- To find a generalized model which performs best in all different domains of fake reviews using comparative analysis of each models performance on different datasets.

II. Related Work

The detection of false reviews has emerged as a key challenge in the field of sentiment analysis as the reliance on internet reviews for product and service evaluation grows. The issue of fake review detection has been the subject of extensive research in recent years, which has used a variety of methods, including machine learning, natural language processing, and network analysis."

In addition to focusing on how to create fake reviews, Salminen J *et al.*'s work [1] also provides techniques for spotting them. The authors want to help merchants and customers understand how critical it is to spot and deal with fake online platform reviews. The researchers created fake reviews using GPT-2 and ULMFit, mixed them with real Amazon product reviews, and used the resulting dataset to identify fake reviews. With this method, fake review detection algorithms can be evaluated and tested under controlled conditions. In order to identify fraudulent reviews in the created dataset, the study used a number of algorithms, including SVM, fake RoBERTa, and the OpenAI fake detection model. The fake RoBERTa model outperformed the other algorithms, according to the data, suggesting that it has the capacity to detect fake reviews with accuracy. The OpenAI model, however, fell short of expectations, underlining the need for more performance enhancement or investigation of alternative strategies for fake review detection. The chance that the Amazon dataset utilised for real reviews may contain false reviews is one potential flaw with the study. This could affect the trustworthiness of the findings and restrict their accuracy. It is crucial to take into account the potential for fraudulent reviews to appear in real-world datasets and to include the necessary methods to deal with this problem in upcoming studies. Another thing to think about is the fake RoBERTa model's inability to recognise human-generated reviews that could evolve over time. The performance of the fake RoBERTa model may deteriorate over time and ongoing updates or fine-tuning may be necessary to maintain its efficacy since human-generated reviews can differ in their traits and attitude.

Using sentiment analysis techniques, Anas S M *et al.* [2] suggest an approach for locating fake reviews and deleting them from a platform. The authors train their models for spotting fake product reviews using datasets from Yelp and Amazon. After training the models with Naive Bayes and Random Forest classifiers, which categorise the reviews as positive, negative, or neutral, the polarity score of the reviews is used to determine if a review is legitimate or fraudulent. The study's conclusions show that the Random Forest classifier outperformed Naive Bayes in terms of accuracy and performance. Because Naive Bayes is a probabilistic model, one of its drawbacks is that it might not effectively capture complicated correlations between features, which could have an impact on how well it detects fake reviews. However, when utilised with categorical data, Random Forest models can be biassed and take a long time to train. Given that they may affect the precision and dependability of the findings, it is crucial to take into account the classifiers' limitations. Future studies can look into different machine learning classifiers or approaches that could be able to get over these restrictions and enhance the effectiveness of fake review detecting tools. Additionally, as the characteristics of false reviews may alter depending on the context, it would be beneficial to further validate the proposed method using a variety of datasets from various platforms and domains. Further insights into the suggested approach's efficacy and resilience in real-world circumstances can be gained by contrasting it with other cutting-edge methodologies and assessing its performance against several evaluation metrics.

In a study by Alsubari S N *et al.* [3], they collected Yelp data and applied the Tf-Idf approach for feature extraction to determine the most relevant features for spotting fake reviews. SVM, Naive Bayes, Random Forest, and AdaBoost were among the machine learning models used to classify the reviews as fake or real, offering a thorough examination of the issue. In comparison to the other models utilised in the study, the study's findings demonstrated that the Random Forest model was the most accurate in spotting fake reviews. Nevertheless, a drawback of models developed for this study is that they might not be able to adapt to changing fake review methods over time, which could cause a reduction in performance as new approaches or patterns of false reviews appear. In order to effectively address this issue, it is crucial to take into account the dynamic nature of fake reviews and the requirement for ongoing monitoring and upgrading of detection algorithms. Future research might concentrate on creating more reliable and adaptable models that can adjust to shifting patterns of fake reviews. For example, techniques like online learning or ensemble methods that can update the model in real-time as new data becomes available could be incorporated. The results of the study are also dependent on Yelp data, thus it would be beneficial to evaluate the suggested method using data from other platforms or domains in order to judge its generalizability and usefulness in various situations. The effectiveness and dependability of the suggested strategy in identifying fake reviews can be further validated by evaluating the performance of the models using various evaluation criteria and contrasting them with other cutting-edge methods.

In order to improve the effectiveness of machine learning models in detecting fake reviews, Budhi G S *et al.* [4] carried out a study in which they used a resampling strategy to equalise the amount of fake and actual reviews in the training data. The study's conclusions demonstrated that the algorithms' capacity to detect fake reviews was enhanced by the use of the resampling technique. Resampling approaches might be helpful when the dataset is unbalanced, with a disproportionately greater proportion of reviews in one class (for example, fake reviews) than in the other class (for example, real reviews). Resampling can balance class distribution and stop models from becoming biassed towards the majority class by oversampling the minority class or undersampling the majority class. It is crucial to remember that the effectiveness of the models can be influenced by the quality of the data and the use of suitable evaluation measures. When the dataset is small, resampling may increase model accuracy, but it also raises the possibility of overfitting since it repeats data from lower classes or removes samples from higher classes, which could result in conclusions that are biassed or erroneous. In order to accurately evaluate the performance of the models, it is crucial to carefully weigh the trade-offs and potential drawbacks of using resampling techniques, such as overfitting and potential information loss. Resampling can be combined with other strategies to improve the effectiveness of machine learning models for detecting fake reviews, including ensemble approaches, feature selection, and synthetic data production techniques.

Using deep neural networks, notably DFFNN (Deep Feed Forward Neural Network) and CNN (Convolutional Neural Network), combined with word embeddings and emotion mining, the paper by Hajek P *et al.* [5] presents a false review detection system. To test the effectiveness of their technique in identifying fake reviews, the researchers trained it using a dataset of Amazon product reviews. The study's findings demonstrated that the proposed strategy was more effective than traditional machine learning models at detecting fake reviews. As opposed to conventional machine learning techniques, this shows that deep neural networks, word embeddings, and emotion mining have the potential to increase the accuracy of fake review detection. However, one drawback of the suggested approach is that it could not be successful in spotting fraudulent reviews created utilising new approaches and techniques. The system may not be able to effectively identify fake reviews that make use of fresh approaches or tactics because it is trained on the present patterns and characteristics of fake reviews. This emphasises the necessity for ongoing upgrades and advancements in fake review detection techniques to stay up with the changing tactics utilised by fake review offenders.

A study on classifying fake reviews using supervised machine learning algorithms was carried out by Khan *et al.* [6]. In order to increase the classifier's accuracy, the study used a variety of feature extraction approaches, such as TF-IDF, Term Frequency, and Sentiment Analysis. According to the study's findings, SVM outperformed the other models in terms of classification accuracy, while K-NN (K-Nearest Neighbours) performed the worst. The generalizability of the findings, it is crucial to highlight, may be difficult due to inherent bias in the dataset against genuine evaluations. Biassed datasets may not accurately reflect the real-world distribution and characteristics of reviews, which can have an impact on the accuracy and dependability of machine learning models. It is critical to take into account the dataset's diversity and representativeness when developing and testing machine learning models. Biassed models may not perform effectively on real-world data as a result of biassed datasets. As a result, it's critical to solve the dataset's bias problem and carefully assess how well the findings can be applied to real-world situations. Additionally, combining several feature extraction methods and investigating other cutting-edge machine learning algorithms may improve the precision and effectiveness of the classification models for detecting fake reviews. To guarantee the dependability and robustness of the results, it is crucial to carry out extensive assessments and validations of the proposed approach utilising a variety of unbiased datasets.

A method of ensemble learning was put out by *et al.* [7] for identifying fake reviews. The study improved the performance of their fake review detection system by using the predictions of various machine learning models. The authors tested their method using a dataset of hotel reviews and used DoC2Vec, a variation of Doc2Vec, for embedding text data. A method called ensemble learning combines the results of various base models to get a single final prediction. Compared to individual models, it frequently leads to increased accuracy and resilience. Incorporating many machine learning models into an ensemble can aid in identifying a variety of patterns and features in the data, improving the accuracy of fake review detection. However, it might be difficult to evaluate ensemble models since the right evaluation metrics may not be the same as those used for individual supervised models. To guarantee that the effectiveness of the ensemble technique is properly examined, it is crucial to carefully consider the assessment metrics and methodology employed in the study. Additionally, comparisons with other supervised models can shed insight into how well the ensemble technique performs in comparison. DoC2Vec is a well-known method for extracting document-level embeddings from text data, making the use of DoC2Vec for text data embedding a novel strategy. The precise properties of the dataset and the quality of the embeddings produced by DoC2Vec may have an impact on this method's effectiveness. To better appreciate the possible advantages and constraints of this technique, more study and evaluation of ensemble learning approaches for fake review identification, including comparisons with other models and evaluation on various datasets, is recommended. In order to determine whether the proposed methodology is effective at detecting fake reviews, it is crucial to make sure that it is extensively validated and that its performance is fairly evaluated.

An ensemble approach for identifying fake internet reviews was put forth by Yao j *et al.* [8]. To increase its accuracy of fake review detection, the model applies feature trimming, parameter optimisation, and data resampling approaches. To address the problem of imbalanced data, the authors used ensemble models, including voting and stacking, and implemented resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) and RUS (Random Undersampling). Feature pruning is a method for narrowing down the initial set of features to the most important ones, which can help to minimise noise and enhance model performance. To get the best performance, parameter optimisation entails fine-tuning the hyperparameters of the machine learning models. In order to overcome the problem of imbalanced data, where the amount of samples in one class is noticeably higher or lower than the other class, which can damage the model's capacity to precisely categorise fake reviews, data resampling techniques, such as SMOTE and RUS, are used. Voting and stacking ensemble models use the results of numerous base models to produce a final prediction. Voting and stacking require training a new model using the predictions of the base models, whereas voting includes merging the predictions of many models using a majority vote. The f1 score, a popular metric for assessing the balance between precision and recall in binary classification problems, revealed that the stacking ensemble model outperformed the voting model. Resampling methods like SMOTE and RUS can be used to address the problem of imbalanced data and enhance the model's effectiveness in spotting false reviews. By lowering noise and fine-tuning the model's hyperparameters, feature trimming and parameter optimisation can also help a model perform better. It is crucial to remember that the dataset's unique properties, the selection of ensemble techniques, and the resampling methods may all affect how effective the suggested ensemble model is. To prove the generalizability and robustness of the suggested approach in detecting fake online reviews, more investigation and validation using a variety of datasets are required. The f1-score of the models rises as the sample size via re-sampling does so because of over-fitting.

The use of an unsupervised LSTM (Long Short-Term Memory) autoencoder for spam review identification is the subject of a study by Saumya S*et al.* [9]. The authors provide a unique method for encoding and decoding reviews using an LSTM autoencoder in order to identify spam reviews unsupervisedly and without the use of labelled data. This paper makes a significant addition by using an unsupervised strategy, which does not require labelled data and may be useful in situations when acquiring labelled data is difficult or expensive. In order to learn representations of reviews and reconstruct them, the LSTM autoencoder is used. This enables the detection of spam reviews based on discrepancies in the reconstructed reviews. It is crucial to keep in mind, nevertheless, that the use of a small dataset or a particular area may have limited the study's conclusions. When used with various datasets or domains, the efficiency of the suggested approach may change, and the results may not be as general. The study also notes that one problem in the field of fake review identification is that the method may not catch spam reviews that are designed to be more subtle or sophisticated.

The Yelp dataset was used in a study by Hassan R *et al.* [10] to identify fake internet reviews using supervised machine learning techniques. In the study, three methods—Logistic Regression, Support Vector Machines (SVM), and Naive Bayes—were used. With accuracy, precision, and recall scores of 0.887, 0.9, and 0.87, respectively, SVM outperformed the others in terms of outcomes. Sentiment Score, Empath Categories, and Term Frequency-Inverse Document Frequency (TF-IDF) were three features the study used to identify fraudulent reviews. Sentiment Score assesses the review's sentiment polarity, Empath Categories depict the text's emotional content, and TF-IDF evaluates the significance of the phrases used. Authors sought to produce a thorough representation of review text that could be utilised for classification by utilising these qualities. It is crucial to remember that the approach's efficacy could be constrained by the calibre and scope of the training dataset. The availability of high-quality, varied, and representative training data is a key factor in determining the accuracy and dependability of machine learning models. The performance and generalizability of the models may be impacted if the training dataset is limited or biassed. To ensure the robustness and reliability of the suggested approach for identifying fake online reviews, significant consideration should be paid to the quality and amount of the training dataset. The accuracy and efficacy of the method may also be increased by utilising bigger and more varied datasets, including domain-specific variables, and researching further advanced machine learning approaches.

For the purpose of identifying fake reviews, Wang J *et al.* [11] provide a technique that integrates a variety of variables, such as social network analysis, sentiment analysis, and emotion analysis. The implementation of a rolling collaborative training technique in this study is noteworthy because it enables the model to continuously update its training data with new data and react to changing patterns in fake review activity over time. The study's conclusions imply that the recommended procedure is superior to previous approaches in detecting fake reviews. The model can be able to capture diverse characteristics of fake reviews and increase its detection effectiveness thanks to the incorporation of various variables and the rolling collaborative training approach. It's important to keep in mind, though, that the quantity of the dataset employed may have an impact on how well this technique performs. The study shows that, in comparison to other deep learning models, the suggested strategy would not perform well with a huge dataset. This emphasises how crucial it is to take into account the dataset's characteristics and the proposed approach's limitations in practical applications. The applicability of this approach to various datasets and domains can be explored in more detail, and it can also be investigated how to improve its performance with big datasets. Further insights into the suggested method's efficacy in fake review detecting tasks can be gained by comparing it to other cutting-edge strategies and performing comparison research.

In their paper, Li J *et al.* [12] offer a technique for detecting fake reviews that entails grouping reviews by subjects and attitudes that are similar, and then using sentiment analysis to find fake reviews within these groups. The authors contend that categorising reviews into groups based on subjects and sentiments with a high degree of similarity can increase the accuracy of fake review detection and act as a guide for sentiment analysis. This study makes a significant contribution by emphasising the importance of taking into account the context of reviews, including attitudes and subjects, as opposed to analysing individual reviews in isolation. The suggested method takes into account the overall patterns and trends within reviews, which may be suggestive of fake reviews, by combining reviews that share similar sentiments and topics. The paper does, however, identify a significant weakness in the suggested methodology, namely the treatment of duplicated reviews as fake reviews. This might not be accurate because duplicate reviews might not necessarily point to phoney reviews. False positives can occur, for instance, when real evaluations from various individuals have identical themes and sentiments. Despite this drawback, the study by Li J *et al.* [12] makes a significant contribution to the body of knowledge on the detection of fake reviews by recommending an approach that accounts for the context of reviews, such as attitudes and subjects, to increase the accuracy of fake review identification. false is unreliable.

The objectives of Ligthart C *et al.*'sstudy [13] is to evaluate semi-supervised learning techniques for fake review detection that have not been thoroughly explored in prior studies. The authors emphasise that when compared to conventional supervised learning techniques, semi-supervised learning techniques may be able to increase the accuracy of fake review detection. The investigation of semi-supervised learning techniques in the context of identifying fraudulent reviews is one important addition of this study. Future research in this area may benefit from this information, which broadens our awareness of other strategies for fake review detection beyond conventional supervised learning techniques. The paper does acknowledge that the calibre and size of the training dataset may have an impact on how well semi-supervised learning systems work. The accuracy and dependability of models created using semi-supervised learning approaches can be impacted by a lack of or poor training data. Despite this drawback, Ligthart C *et al.*'sstudy [13] sheds light on the potential of semi-supervised learning techniques for fake review detection and suggests that they may be useful in enhancing the precision of false review identification. By examining various semi-supervised learning techniques, enhancing their performance in relation to the calibre and quantity of training data, and assessing their efficacy on various datasets and scenarios, additional research can build on this work to better understand their applicability in real-world fake review detection.

A novel multi-iterative graph-based model is presented by Noekhah S *et al.* [14] and captures the interaction between reviews and users. The model is trained using content-based, behavior-based, and relation-based characteristics, and iteratively updates the graph representation of the reviews.This study makes a significant addition by combining many traits, such as content-based, behavior-based, and relation-based features, which led to better performance than a single model. This underlines the significance of taking into account a variety of variables in fake review detection as they can compliment one another and improve accuracy and precision. Furthermore, the study discovered that in terms of accuracy and precision, the suggested multi-iterative graph-based model surpassed current state-of-the-art fake review identification techniques. This shows that the unique methodology put out by Noekhah S *et al.* [14] has the potential to be more exact and effective than current techniques at spotting fake reviews. It's important to remember that the quality and amount of the data, the unique aspects of the dataset, and the selection of features and parameters can all have an impact on how well the suggested model performs. To better understand the generalizability and application of the multi-iterative graph-based model in real-world fake review detection tasks, additional study can explore these parameters and evaluate its efficacy on other datasets and scenarios.

For the purpose of identifying fake reviews, Liu *et al.* [15] suggest a bidirectional LSTM (BiLSTM) model with feature representation. Word embeddings, part-of-speech tags, sentiment scores, and other features are taken from the review text and fed into the BiLSTM model. The incorporation of several attributes in addition to the review text, which can provide additional contextual information for false review detection, is one noteworthy component of this study. The suggested BiLSTM model may be able to capture multiple characteristics of fake reviews by adding a variety of variables, which could enhance accuracy, F1-score, and precision in comparison to current approaches. The results of the experiment on a benchmark dataset indicate that the suggested methodology performs better in terms of accuracy, F1-score, and precision than various other existing methods. This suggests that the suggested BiLSTM model with feature representation may be useful for identifying fake reviews. However, it's vital to take into account the study's limitations, such as the unique traits of the benchmark dataset utilised, the likelihood of bias in feature selection, and the applicability of the suggested strategy to various datasets and domains. To better understand how the BiLSTM model with feature representation performs on various datasets and how it applies to real-world fake review detection tasks, additional study can look into these variables.

III. PROPOSED METHODOLOGY

The first step in conducting a comprehensive analysis of fake reviews is to collect relevant datasets containing fake review data. These datasets are available from a variety of sources, including online review websites and publicly available research datasets. After gathering the datasets, the data must be cleaned by removing any unwanted rows or irrelevant information to ensure that the analysis is based on accurate and relevant data.

Following data cleaning, the datasets go through data preprocessing, which consists of several steps. First, any special characters, numbers, or other irrelevant information are removed from the reviews. Then, to reduce noise in the data, stop words, which are common words with little meaning, are removed. To normalise the text data, the reviews are then lemmatized, which involves reducing words to their root form. Finally, the reviews are tokenized, which means they are broken down into individual words or phrases, and embedding techniques like GloVe are used to convert the text data into numerical representations for further analysis.

Following the pre-processing of the data, various machine learning models are trained on the datasets. These models include, among others, Bi-directional Long Short-Term Memory (Bi-LSTM), DistilBERT, Xlnet, ALBERT, DeBERTa, and fine-tuned DistilBERT. Depending on the model architecture and dataset size, these models are trained on pre-processed datasets using appropriate techniques such as supervised learning or transfer learning.

The models are evaluated on various test datasets after they have been trained to assess their performance. Performance indicators including recall, accuracy, and F1 score are used to measure how well the models perform at spotting fake reviews.

The performance of several models is then compared in a thorough performance analysis. This analysis entails assessing the benefits and drawbacks of each model, pinpointing their shortcomings, and comprehending how they work in various settings. This analysis aids in the formulation of conclusions and suggestions for the efficient application of these models for fake review detection.

IV. Datasets and Preprocessing

A. Datasets

1) Fake review dataset, by Joni Salminen et al: The 40,000 reviews in the fake review dataset, developed by Joni Salminen

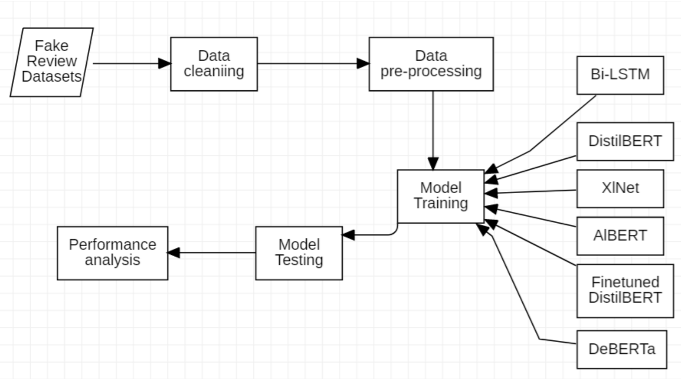

Fig. 1. Proposed Methodology Diagram

et al. [1], are evenly divided between real Amazon reviews and artificially generated fake reviews. The fake reviews were generated using GPT-2 and ULMFit, which gives the dataset a special quality because it contains fake reviews that were produced artificially. This dataset offers the chance to assess the effectiveness of fake review detection models in a controlled setting where the false reviews are produced using particular methods. It as a useful tool to compare the performance of various models and gauge how well they are able to identify fake reviews produced by language models.

2) Ott et.al dataset: Another important dataset used in fake review detection studies is the Ott et al. dataset. This dataset consists of 1600 hotel reviews for 20 hotels in Chicago, evenly distributed between reviews that are genuinely positive, deceptively positive, genuinely negative, and falsely negative. The dataset includes reviews from multiple websites, including Yelp, Expedia, Hotels.com, Mechanical Turk, Orbitz, Priceline, and TripAdvisor, offering a variety of sources and viewpoints. The dataset gains realism through the usage of Mechanical Turk to generate fake reviews because it accurately depicts the existence of fake reviews on real online review sites. With varied levels of emotion and authenticity, this dataset enables researchers to assess the effectiveness of fake review detection methods in a more realistic setting.

3) Amazon fake review dataset: Amazon fake review dataset: A common dataset used in false review detection research is the Amazon fake review dataset, which is accessible on Kaggle. There are 21,000 product reviews total, with 10,500 real and 10,500 fake ones distributed equally. The dataset offers a fair representation of attitudes by incorporating both fake and real assessments. In the context of e-commerce and online retail, where fake reviews are common and can influence consumer decision-making, this dataset is pertinent for assessing the effectiveness of fake review detection programmes. This dataset's accessibility to a sizable number of reviewers enables thorough evaluations of the effectiveness of various models and their applicability to actual-world situations.

So therefore, each of these datasets—the Ott et al. dataset, the Amazon fake review dataset, and the Fake review dataset by Joni Salminen et al.—offer useful resources for assessing the effectiveness of fake review detection programmes. They provide various viewpoints and traits, such as artificially generated fake reviews, realistically produced false reviews via Mechanical Turk, and a combination of favourable and unfavourable evaluations from online shopping platforms. These datasets can be used to evaluate and compare various algorithms, comprehend their strengths and limitations, and pinpoint areas in which fake review detection research needs to be improved.

TABLE I
DATASET DISCRIPTIONS

Datasets	**Size**	**No of Fake Reviews**	**No.of Real** Reviews
Fake review by Joni [1]	40,000	20,000	20,000
Amazon fake review	21,000	10,500	10,500
Ott et.al dataset	1,600	800	800

B. Preprocessing

The reviews underwent a process of ”decontraction,” which involved converting contractions such as ”won't” and ”can't” into their full form. Additionally, common words known as ”stop words” were removed because they appear frequently in reviews but do not provide meaningful information. Examples of stop words include ”the,” ”is,” and ”was.” The preprocessed reviews were also subjected to tokenization and lemmatization to further refine their content. Lemmatization reduces the tokens to their simplest form, whereas tokenization divides the text into smaller parts known as tokens. This enables a clearer and more informative representation of the reviews.

V. IMPLEMENTATION

A. Models

1) Bi-LSTM: : Bidirectional Long Short-Term Memory (Bi-LSTM) is a type of recurrent neural network (RNN) architecture that has become widely used in natural language processing (NLP) applications. The classic LSTM model has been modified to enable bidirectional processing of input sequences. As a result, the network may collect both the current input and the context in both directions while processing the sequence both forward and backward simultaneously. In this experiment Bi-LSTM model was trained for 30 epochs with batch size of 64.

2) DistilBERT: : DistilBERT is a state-of-the-art language model, It is a condensed version of the BERT model that is intended to be quicker and use less memory while still retaining a high degree of accuracy in various NLP tasks. In order to minimise the number of parameters in the BERT model without sacrificing its effectiveness for tasks like text categorization, sentiment analysis, and question-answering, DistilBERT uses a distillation method. In this experiment DistilBERT model was trained for 6 epochs with learning rate 2e-5.

3) DeBERTa: : It is an extension of the original BERT model that includes various cutting-edge features, including cross-layer parameter sharing, large-scale training, and dynamic masking. Instead of utilising a fixed mask, as in the original BERT, the dynamic masking technique enables the model to learn to mask various portions of the input sequence during training. It has been demonstrated that the model performs better on a variety of downstream tasks when large-scale training is used, which entails training the model on enormous volumes of data. The model's efficiency and performance are further enhanced by the cross-layer parameter sharing technique, which allows the model to share parameters across many levels. In this experiment DeBERTa model was trained for 3 epoch with a batch size of 16 and learning rate 2e-5.

4) AlBERT: : It is intended to address some of the original BERT model's shortcomings, such as its size and long training time. ALBERT accomplishes this by training the model on longer sequences and employing a parameter-sharing approach known as ”factorization.” The factorization technique makes the model smaller and faster while reducing the number of parameters and keeping its ability to represent data. ALBERT also adds a self-supervised loss function, which motivates the model to gain more knowledge from longer sequences. In this experiment AlBERT model was trained for 6 epoch and learning rate 2e-5.

5) XlNET: : Unlike BERT, which utilises a masked language modelling aim, XLNet uses a permutation-based training strategy that lets it to consider all potential permutations of the input sequence during training, thereby overcoming the restrictions of the left-to-right or right-to-left pre-training allows it to better capture the dependencies between all words in a sentence. In this experiment Xlnet model was trained for 3 epoch with batch size 16 and learning rate 2e-5.

6) Fine-tuned DistilBERT: : In order to fine-tune the DistilBERT model, we froze all of its layers, attached two of our own neural network layers, and trained this new model. Keep in mind that the model's pre-trained weights don't change, and the error only back-propagates through the attached layers. During model training using the new dataset, the weights of only the attached layers will be modified. In this experiment the fine-tuned model was trained for 6 epochs and learning rate 2e-5.

VI. Results

A. Metrics

Metrics are essential for assessing how well fake review detection techniques are working. Different measures are frequently employed to rate the effectiveness of various techniques. A frequently used metric for evaluating how accurately a model's predictions are generally made is accuracy. Precision, which measures the accuracy of the model's ability to distinguish between real and fake reviews, is measured as the ratio of true positive predictions to all positive predictions. The ability of the model to identify all true positive cases is measured by recall, sometimes referred to as sensitivity or true positive rate. The harmonic mean of precision and recall, or F1-score, offers a balanced evaluation of the model's performance. These metrics enable evaluation and comparison of the efficacy of various methods and offer insightful data on the overall performance of fake review detecting techniques.

Accuracy:

$$\frac{TP+TN}{TP+FP+TN+FN}$$

Precision:

$$\frac{TP}{TP+FP}$$

Recall:

$$\frac{TP}{TP+FN}$$

F1-Score:

$$2*\frac{Precision*Recall}{Precision+Recall}$$

B. EXPERIMENT ON OTT ET.AL DATASET

The outcomes on ott et.al dataset from our tests demonstrate that different models perform differently when it comes to detecting fake reviews. AlBERT had the best accuracy of the models tested, at 0.89; it also had the highest precision, recall, and F1-score, at 0.89, 0.88, and 0.89, respectively. Achieving an accuracy of 0.88, XlNET and DistilBERT both performed well. Bi-LSTM and DeBERTa both attained accuracy values of 0.76, but fine-tuned DistilBERT had a lower accuracy of 0.65.

AlBERT, XlNET, and DistilBERT consistently displayed high precision scores of 0.89, 0.88, and 0.87, demonstrating the capability of these models to correctly identify fake reviews. DeBERTa had a precision score of 0.79 while finetuned DistilBERT and Bi-LSTM had somewhat lower precision scores of 0.66 and 0.76, respectively.

The recall scores for the majority of the models were also reasonably high, ranging from 0.85 to 0.88, showing that these models are capable of accurately identifying a significant portion of actual fake reviews. With a recall score of 0.66, Finetuned DistilBERT, on the other hand, may have overlooked some fake reviews in the dataset.

AlBERT received the highest F1 score of 0.89, followed by XlNET with an F1 score of 0.87. The F1 scores for the various models varied. Finetuned DistilBERT and DeBERTa had F1-scores of 0.65 and 0.76, compared to 0.85 and 0.76 for DistilBERT and Bi-LSTM, respectively.

TABLE II
Results on OTT ET.AL DATASET

Model	Accuracy	Precision	Recall	F1-score
Bi-LSTM	0.76	0.76	0.76	0.76
DistilBERT	0.85	0.87	0.85	0.85
Finetuned DistilBERT	0.65	0.66	0.66	0.65
XlNET	0.88	0.88	0.87	0.87
AlBERT	0.89	0.89	0.88	0.89
DeBERTa	0.76	0.79	0.77	0.76

C. EXPERIMENT ON AMAZON FAKE REVIEW DATASET

Based on the results of our experiments Amazon fake review dataset, it can be seen that different machine learning models performed differently in terms of identifying fake reviews. With values of 0.59 for all parameters, Bi-LSTM had the lowest accuracy, precision, recall, and F1-score. The accuracy scores of DistilBERT, Finetuned DistilBERT, XlNET, and AlBERT were all 0.67, but DeBERTa did somewhat better with an accuracy of 0.72.

DeBERTa received the highest precision score, 0.72, while XlNET received a precision score of 0.69. AlBERT, DistilBERT, and Finetuned DistilBERT all had precision scores of 0.68, while Bi-LSTM's precision score was the lowest at 0.59.

The models' recall scores, which ranged from 0.67 to 0.72, were all consistent, showing that they were all able to properly detect a similar percentage of genuine false reviews.

The F1-scores ranged from 0.66 to 0.72 for the majority of the models, with DeBERTa receiving the highest F1-score of 0.72.

TABLE III
Results on Amazon Fake Review Dataset

Model	Accuracy	Precision	Recall	F1-score
Bi-LSTM	0.59	0.59	0.59	0.59
DistilBERT	0.67	0.68	0.67	0.67
Finetuned DistilBERT	0.67	0.68	0.67	0.67
XlNET	0.67	0.69	0.67	0.66
AlBERT	0.67	0.67	0.67	0.67
DeBERTa	0.72	0.72	0.72	0.72

D. EXPERIMENT ON FAKE REVIEW DATASET, BY JONI SALMINEN ET.AL

Based on the outcomes of our tests on fake review dataset, by joni salminen et.al, it is clear that the machine learning models were highly effective in spotting false reviews. DistilBERT, AlBERT, and XlNET all had accuracy scores of 0.97, which indicates that they were highly accurate in classifying fake reviews. Both Bi-LSTM and DeBERTa received scores for accuracy that were comparatively high, 0.88 and 0.96, respectively.

The models' precision scores, which ranged from 0.86 to 0.97, were consistently high, showing that they could accurately identify fake reviews with a low percentage of false positives. The models were able to correctly detect a significant portion of actual fake reviews, as seen by the recall scores, which ranged from 0.94 to 0.97.

The F1-scores ranged from 0.85 to 0.97 for the majority of the models, with AlBERT receiving the highest F1-score of 0.97. This suggests that the models were successful in striking a solid balance between precision and recall, leading to the precise and trustworthy detection of false reviews.

TABLE IV
Results on FAKE REVIEW DATASET, BY JONI SALMINEN ET.AL

Model	Accuracy	Precision	Recall	F1-score
Bi-LSTM	0.76	0.76	0.76	0.76
DistilBERT	0.85	0.87	0.85	0.85
Finetuned DistilBERT	0.65	0.66	0.66	0.65
XlNET	0.88	0.88	0.87	0.87
AlBERT	0.89	0.89	0.88	0.89
DeBERTa	0.76	0.79	0.77	0.76

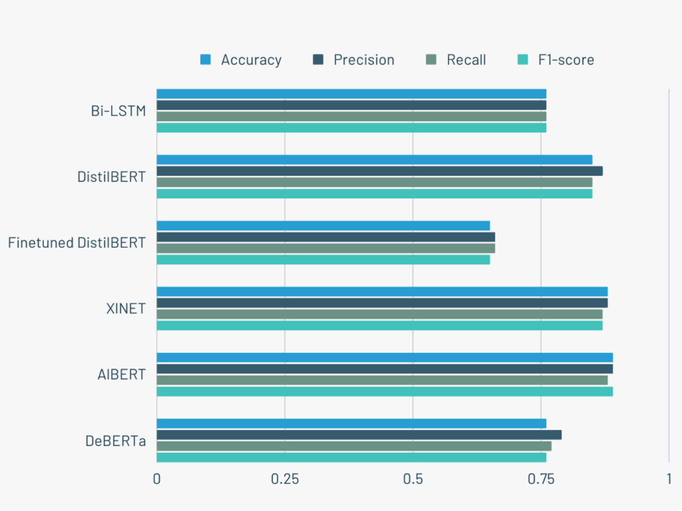

Fig. 2. Bar Graph Of Performances Of Different Mondel On Ott ET.AL

E. DISCUSSION

VII. CONCLUSION AND FUTURE WORK

AlBERT consistently displayed strong performance across all assessed measures, including accuracy, precision, recall, and F1-score, as seen by the bar graph (Fig 2). It received the highest accuracy score of 0.89, indicating that 89% of the evaluations in the OTT dataset were properly categorised. AlBERT also received the highest precision score of 0.89, indicating that it had the lowest false positive rate and correctly detected actual fake reviews without misclassifying fake reviews as real.

With scores for accuracy of 0.88 and 0.85 in the OTT dataset, XlNET and DistilBERT also did well. The precision ratings for both models were similar, with XlNET marginally surpassing DistilBERT. As a result, XlNET and DistilBERT are likely choices for false review identification in the OTT dataset because their ability to properly and precisely classify reviews with a high degree of precision is demonstrated by this.

With accuracy, precision, recall, and F1-score values of 0.76, respectively, Bi-LSTM and DeBERTa performed similarly. They displayed a respectable degree of performance in the OTT dataset, even though they outperformed AlBERT, XlNET, and DistilBERT.

On the other hand, Finetuned DistilBERT performed worse than other models, scoring 0.65 for accuracy, precision, recall, and F1-score. This shows that additional experimentation and fine-tuning may be required to enhance its performance in the OTT sample.

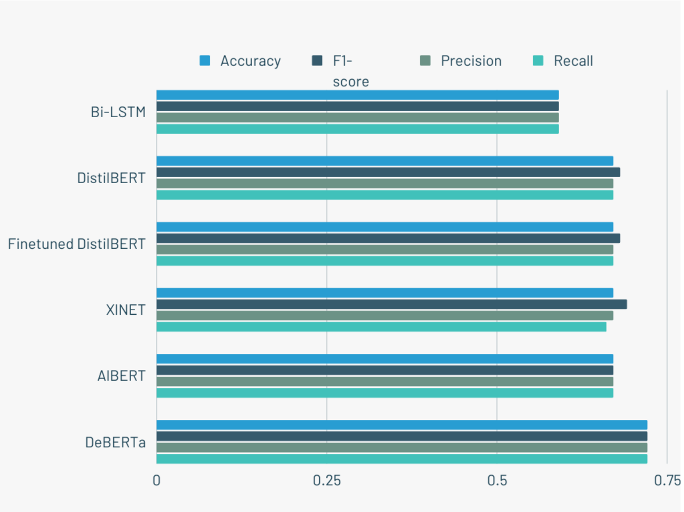

Fig. 3. Bar Graph Of Performances Of Different Mondel On Amazon Fake Review Dataset

The Fig 3 showing the experimental outcomes for detecting fraudulent reviews in the Amazon dataset demonstrates that DeBERTa outperformed all other methods in terms of accuracy, precision, recall, and F1-score. It received an accuracy score of 0.72, meaning that 72% of the reviews in the Amazon dataset were correctly identified. DeBERTa also received the best precision score (0.72), which indicates that it had the lowest false positive rate and correctly identified true positive reviews without misclassifying fake reviews as real.

With accuracy scores ranging from 0.67 to 0.68, DistilBERT, Finetuned DistilBERT, XlNET, and AlBERT all performed similarly to one another. XlNET significantly outperformed the other models, but they all had comparable precision ratings. This shows that DistilBERT, Finetuned DistilBERT, XlNET, and AlBERT were able to correctly categorise reviews in the Amazon dataset with a respectable degree of precision.

Bi-LSTM performed the worst out of the models, scoring 0.59 for accuracy, precision, recall, and F1-score. This suggests that additional development could be needed to increase its performance in the Amazon dataset.

The top-performing models for fake review identification are DistilBERT, AlBERT, and XlNET, from the bar graph (Fig 4) according to experimental results on the Fake Review Dataset by Joni Salminen et al. These models consistently achieve excellent accuracy, precision, recall, and F1-score values. With an accuracy of 0.97 and consistently strong ratings for other measures, DistilBERT outperformed all other models. With accuracy ratings of 0.97 and 0.94, respectively, AlBERT and XlNET likewise fared remarkably well. According to these results, transformer-based models like DistilBERT, AlBERT, and XlNET are very good at spotting fake reviews in the dataset created by Joni Salminen et al.

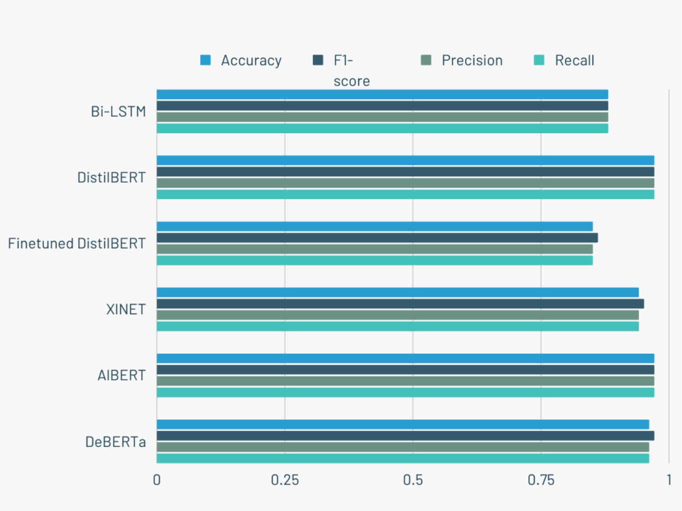

Fig. 4. Bar Graph Of Performances Of Different Mondel On Fake Review Dataset, By Joni Salminen Et.Al

Finetuned DistilBERT and DeBERTa that had been improved performed well as well, with accuracy scores of 0.85 and 0.96, respectively. The top three models, however, outperformed them somewhat in terms of accuracy and other criteria. Bi-LSTM, on the other hand, had a significantly worse performance, scoring 0.88 for accuracy.

These results suggest that the Fake Review Dataset by Joni Salminen et al is well suited for fake review identification by transformer-based models, particularly DistilBERT, AlBERT, and XlNET. These models have excellent accuracy and other metric scores, pointing to their potential for usage in practical applications in the real world. To improve the models' performance and find the best method for fake review detection in this dataset, more testing, tweaking, and optimisation may be required

Overall, the performance of different models can be compared based on the experimental findings from the three datasets (OTT, Amazon, and Fake Review Dataset by Joni Salminen et al.).

With high accuracy scores of 0.85-0.97, DistilBERT and AlBERT consistently outperformed the competition across all three datasets. XlNET also performed well, with accuracy ratings ranging from 0.67 to 0.94. Finetuned DistilBERT and DeBERTa performed moderately, with accuracy scores ranging from 0.65 to 0.96. The accuracy scores for Bi-LSTM, on the other hand, ranged from 0.59 to 0.88.

Similar patterns were seen in other metrics, including as precision, recall, and F1-score, with DistilBERT, AlBERT, and XlNET consistently outperforming other models across datasets. Moderate performance was also shown by tweaked DistilBERT and DeBERTa, although Bi-LSTM performed significantly worse in terms of these parameters.

VIII. CONCLUSION AND FUTURE WORK

As a conclusion, transformer-based models like DistilBERT, AlBERT, and XlNET are viable contenders for fake review detection tasks, according to the experimental results from the three datasets (OTT, Amazon, and fake Review Dataset by Joni Salminen et al.). High accuracy, precision, recall, and F1-score were repeatedly displayed by these models, indicating their potential for usage in real-world situations involving the detection of false reviews in many domains. It is crucial to remember that the performance of each model may differ based on the precise dataset and job, and that additional fine-tuning and experimentation may be required to find the model that performs best for a certain dataset. Future scope of the work for improvement may also be included.

Future research may also examine various methods for enhancing and optimising the performance of Bi-LSTM, which demonstrated relatively poorer performance in comparison to transformer-based models. To increase the accuracy and other metrics of Bi-LSTM in fake review detection tasks, this can entail adding more features, investigating various topologies, or using ensemble approaches. The interpretability and explainability of the transformer-based models could also be the subject of future research, as doing so can help us better understand how these models make decisions and develop confidence in their forecasts.

While transformer-based models exhibit promise in fake review detection tasks, more investigation and testing are required to improve model performance, understand model limitations, and look into ways to make these models simpler to understand and interpretable for practical applications.

REFERENCES

[1] Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771.

[2] Anas, S. M., & Kumari, S. (2021, January). Opinion mining based fake product review monitoring and removal system. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 985-988). IEEE.

[3] Alsubari, S. N., Deshmukh, S. N., Alqarni, A. A., Alsharif, N., Aldhyani, T. H., Alsaade, F. W., & Khalaf, O. I. (2022). Data analytics for the identification of fake reviews using supervised learning. CMC-Computers, Materials & Continua, 70(2), 3189-3204.

[4] Budhi, G. S., Chiong, R., & Wang, Z. (2021). Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. Multimedia Tools and Applications, 80(9), 13079-13097.

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

[6] Khan, H., Asghar, M. U., Asghar, M. Z., Srivastava, G., Maddikunta, P. K. R., & Gadekallu, T. R. (2021, January). Fake review classification using supervised machine learning. In International Conference on Pattern Recognition (pp. 269-288). Springer, Cham.

[7] Gutierrez-Espinoza, L., Abri, F., Namin, A. S., Jones, K. S., & Sears, D. R. (2020, July). Ensemble learning for detecting fake reviews. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 1320-1325). IEEE.

[8] Yao, J., Zheng, Y., & Jiang, H. (2021). An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization. IEEE Access, 9, 16914-16927.

[9] Saumya, S., & Singh, J. P. (2020). Spam review detection using LSTM autoencoder: an unsupervised approach. Electronic Commerce Research, 1-21.

[10] Hassan, R., & Islam, M. R. (2020, December). A Supervised Machine Learning Approach to Detect Fake Online Reviews. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[11] Wang, J., Kan, H., Meng, F., Mu, Q., Shi, G., & Xiao, X. (2020). Fake review detection based on multiple feature fusion and rolling collaborative training. IEEE Access, 8, 182625-182639.

[12] Li, J., Lv, P., Xiao, W., Yang, L., & Zhang, P. (2021). Exploring groups of opinion spam using sentiment analysis guided by nominated topics. Expert Systems with Applications, 171, 114585.

[13] A. Ligthart, C. Catal, and B. Tekinerdogan, “Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification,” Appl. Soft Comput., vol. 101, Mar. 2021, Art. no. 107023.

[14] Noekhah, S., binti Salim, N.,& Zakaria, N. H. (2020). Opinion spam detection: Using multi-iterative graph-based model. Information Processing & Management, 57(1), 102140.

[15] Liu, W., Jing, W., & Li, Y. (2020). Incorporating feature representation into BiLSTM for deceptive review detection. Computing, 102(3), 701-715.