Combination between Deep Learning and Transformer Models to Detect Fake Yelp Electronic Product Reviews

Maysara Mazin Badr Alsaad[1*], Hiren Joshi[2]

[1]Research Scholar, Department of Computer Science, Gujarat University, Ahmedabad, Gujarat, India, Email: maysara@gujaratuniversity.ac.in
[2]Prof, Department of Computer Science, Rollwala Computer Centre, Gujarat University, Ahmedabad, Gujarat, India.
*Corresponding Author

Received: 13.07.2024 Revised: 16.08.2024 Accepted: 15.09.2024

ABSTRACT
Online product reviews play a major role in the success or failure of an E-commerce business. Before procuring products or services, the shoppers usually go through the online reviews posted by previous customers to get recommendations of the details of products and make purchasing decisions. There has been a rise in illusive review spam, which are fake reviews that are designed to appear genuine. Fake, strident, spam, misleading reviews are those written by those who do not have personal experiences with the topics of the reviews. Spammers spread fake reviews in order to denigrate or promote a specific brand or product, persuading consumers to purchase from that brand or not. The detection of genuine ratings and ratings-based reviews across the entire online platform, particularly Yelp product datasets, is the secondary objective. For the purpose of identifying fake online reviews in the e-commerce industry, the paper makes a number of novel hybrid techniques (transformer-based & deep learning), including BiLSTM-CNN, BERT-CNN and RoBERTa-CNN. According to the trial results, the BiLSTM-CNN procedure productively identifies counterfeit internet based audits with a high accuracy of 90% whereas other hybrid models also perform competitively. Moreover, it BiLSTM-CNN model exhibits the most favorable combination of training and testing times among the evaluated models.

Keywords: e-commerce, Fake Reviews, Spam Detection, Deep Learning, Transformer, BiLSTM, CNN, BERT, RoBERTa

1. INTRODUCTION

Individuals and businesses are progressively utilizing reviews online to assist them make purchase decisions as well as how to operate a business. For firms and individuals, high ratings can result in massive financial gains and fame. Regrettably, this presents pretenders with tremendous reasons to exploit the system by posting false ratings either to praise or degrade specific target items or firms. These groups are regarded as opinion spammers & their operations are recognized as opinion spamming. The problem of spam or fraudulent ratings has risen in recent years, and lots of high incidents have been reported in public. Even consumer sites have accumulated a huge amount of automated fraud detection suggestions. Furthermore, there have already been media inquiries where fake reviewers have publicly stated accepting funds to make artificial reviews [1].

There are already several user reviews published for a vast range of goods and services thanks to the swift growth of internet retailing. A significant pool of potential visitors relies on them to analyze the caliber of products or services prior making the purchase. As an outcome, based on a desire of gain or competition, businesses and sellers develop motives and practices to alter reviews, deliberately publishing falsified feedback to purposely deceive potential consumers and manipulate their risky purchasing decisions. An individual (known as an individual spammer) or a gang (known as a spammer group) may also be sponsored by makers in order to publish enhanced positive perceptions on their items or damaging negative reviews on that of their counterparts in order to enhance customer satisfaction and brand [2].

With the growth of online information today, people tend to see reviews first for the places they want to visit, such as restaurants, hotels, or other businesses they need or before they go and buy some product. Yelp is an advertising service and a forum for audience review, which individuals normally utilise to post some review about their business views. Statistics show that by the end of 2018, there have been more than 177 million reviews on the Yelp website. It is benefiting both consumers and businesses. For a business owner, they get free advertising from people who give a useful and positive review of their

business. Unfortuantely, the problem arises when a small portion of irresponsible business owners try to boost up their market by hiring people to create some fake reviews about their business on Yelp website.
Two forms of spammy opinions could be identified, often referred to as fake reviews. Customers who write bad reviews harm the organizations. Positive reviews to encourage manufactured goods/enterprises to have untruthful thoughts. These evaluations are usually fake or dishonest reviews because they are difficult to distinguish by the readout [3].

2. LITERATURE REVIEW

In the past literature, many deep learning techniques and hybrid deep learning models have been used for spam detection of Yelp product reviews.
[4] have presented research work to inspects the influences of social interaction of reviewers' deception recognition at online customer reviews, in their experiment, Yelp's product reviews dataset was gathered and preprocessed. Then they mined behavioral and social relations features of customers. For classification, the authors applied the back propagation neural network classifier for performing the classification of the review text into truthful or fake.
[5] have proposed framework of significant features for deceptive review detection. Based on online Yelp product reviews, they carried out experiments using different supervised machine learning techniques. In terms of features, reviewer (personal, social, review activity, and trust) and review features (sentiment score) were used.
[6] Users increasingly rely on crowd sourced information, such as reviews on Yelp and Amazon, and liked post sand ads on Facebook. This has lento market for black hat promotion techniques via fake (e.g., Sybil) and compromised accounts, and collusion networks. Existing approaches to detect such behavior relies mostly on supervised (or semi-supervised) learning over known (or hypothesized) attacks. They are unable to detect attacks missed by the operator while labeling, or when the attacker changes strategy.
In [7], worked on restaurant reviews that are identified by Yelp's filtering algorithm as suspicious, or fake. They found that nearly one out of five reviews is marked as fake by Yelp's Algorithm. These reviews tend to be more extreme than other reviews and are written by reviewers with less established reputations. Moreover, their finding suggests that economic incentives factor heavily into the decision to commit fraud. Organizations are more likely to game the system when they are facing increased competition and when they have poor or less established reputations.
In [8], dove down to Yelp's secret filtering algorithm. They put a few existing research methods to the test and evaluated performance on the real-life Yelp data. They found the behavioral features perform very well, but the linguistic features are not as effective. Their analysis and experimental results shows that Yelp's filtering is reasonable and its filtering algorithm seems to be correlated with abnormal spamming behaviors.
In [9], They used a large set of reviews from Yelp restaurants and its filtered reviews to characterize the way opinion spamming operates in a commercial setting. Using time-series analysis, they found that there exist 10 three dominant spamming policies: early, mid and late across the various restaurant. Their analysis showed that the deception rating time-series for each restaurant had statistically significant correlations with the dynamics of truthful rating time-series indicating that spam injection may potentially be coordinated by the restaurants/spammers to counter the effect of unfavorable rating over time.
In [10] paper, an innovative framework, titled (Net Spam) use spam feature aimed at showing analysis datasets as mixed material associations near configuration spam area technique hooked on plan issue associations. Using the significance of spam features secure well again achieves terms of special estimations test genuine overview datasets from Yelp and Amazon destinations. inspection that maped load using meiosis ideac potent identifying junk reviews and leads to a good performance.
In [11] paper, because it is necessary to differentiate between fake and genuine reviews, binary classification has become a significant challenge. The Hybrid (LSTM+CNN) model employing Camem-BERT accomplished the most excellent performance in classifying the online reviews of French with 93% accuracy.
The CNN-LSTM is a model proposed by [12], this model has been applied to the different standard fake review datasets for analysing the fake reviews in the in-domain and cross-domain. The LSTM model is combined with CNN to analyze the contextual information from the texts.
The drawback of using deep learning models is they can be used for large datasets and the doesn't provide any computation parallelly. To overcome this disadvantage the transformer models have been introduced. Transformer models are also considered pre-trained models which are already with a huge dataset and also work in analysing the texts in both directions. Some of the transformer models like BERT. RoBERTa and DISTILBERT have been previously used in analyzing fake reviews. Among them, the RoBERTa model

performed well [13].
In [14] six different machine learning algorithms have been reviewed for the given problem statement. It has been concluded that the sentiment analysis method artificial intelligence and communication technologies 657 provides one of the most accurate results when it comes to sorting spam reviews in websites. The paper combines three of the best result producing models namely, GRNN, BI-LSTM and LSTM.
In[15] paper, a comparative analysis of BERT, Hybrid fastText-BiLSTM, and fastText Trigram models to address challenges in achieving more precise sentiment predictions of fake reviews has been performed. Introducing fine-tuned BERT and Hybrid fastText-BiLSTM models for extensive datasets, the study demonstrates that the proposed fine-tuned BERT model outperforms other deep learning models and gives a 0.91 accuracy result.
In[16] paper, effectively identified helpful reviews by employing three distinct approaches: a supervised approach (Fasttext, SVM, Bi-LSTM, CNN, RCNN), a semi-supervised approach (RCNN), and a pre-trained model approach (BERT and RoBERTa), using an product reviews dataset across four domains. Their comparative analysis revealed that among all the approaches, the RCNN model demonstrated superior performance.
Furthermore, we have collected a second highlighted batch of previous research using the Yelp and other dataset using deep learning techniques, which has been arranged and highlighted as shown in the following Table 1.

Table 1. Comparison of standard study approach outcomes using deep learning methods

Ref	Dataset	Methods & Features	Results	Comments
Li et al. [17]	Hotel, Restaurant, and Doctor.	Sentence Weight Neural Network & Word2vec (Skip- gram).	Accuracy: 79.5% Precision:76.1% Recall:89.9% F1:82.3%	CNN outperformed LSTM in a mixed-domain comparison.
Zhao et al. [18]	AMT	Word Order-Preserving CNN & Word2vec and word order.	CNN accuracy:70.02%	CNN cannot handle lengthy articles. So, use the hand-annotated method, which is labor-intensive.
Wang et al. [19]	Yelp Chi dataset	Unsupervised neural network model & Word2vec (CBOW). Behavioral Features.	Accuracy: Hotel:65.4% Restaurant:62%	Learning review embedding with encodes behavioral and linguistic features is effective.
Zhang et al. [20]	AMT dataset Deceptive dataset	DRI-RCNN, Word2vec & Skip-gram.	Accuracy AMT t:82.9% Misleading:80.8%	For whatever reason, this model neglected to account for the behavioral aspects that could enhance efficiency.
Li et al. [21]	Yelp NYC Yelp Zip	CNN & Glove algorithm.	F1-measure:85% for regular reviews and 27% for fake reviews.	They discovered that the quality of the customer's social connections substantially impacted classification accuracy.
Yuan et al. [22]	Yelp NYC Yelp Zip	Unsupervised model & Extracted Real behavior features	F1 measure on Hotel:60% & Restaurant :70%	The importance of link re-weighting in improving performance.
Cao et al. [23]	Hotel reviews from Trip Advisor	LOF algorithm, Aspectrating and TF-IDF.	Accuracy: 79.6% Precision:79% Recall:80.7% F1-score:79.8	Aspect rating performed nicely. Fake review identification could be extended by incorporating additional features. Uses a small number of datasets.

3. METHODOLOGY

The structure of the proposed approach in this research. There are five steps in it. The basic layout of the proposed detection framework is explained in Figure 1.

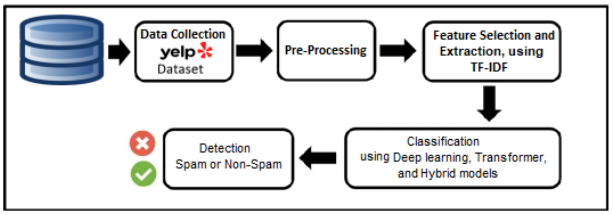

Figure 1. Framework for the proposed methodology

3.1. Data collection

To design and build an efficient and robust fake review detecting model for e-commerce product reviews using a popular types of machine learning models which follows the idea of modelling the probability for classifying fake and genuine reviews. Also provide a consensus strategy for feature extraction and text preprocessing.

The dataset used in this work is consists of 30,476 reviews distributed as 15,473 Trust reviews and 15,003 fake reviews related to different electronic products gathered from the Yelp e-commerce website by [24]. The resulting corpus was composed of reviews from four important cities in the USA: New York, San Francisco, Los Angeles, and Miami. Yelp offers the possibility of searching for category businesses in each city, so the scraping process is focused on the pages on which all electronic businesses from the different selected cities appeared as described in the below Table 2.

Table 2. The size of reviews per USA city

City Name	Fake Reviews	Trusted Reviews
Los Angeles	6270	6009
Miami	1696	1767
New York	3865	3979
San Francisco	3642	3248

3.2. Data preprocessing

Dataset' and the unlabeled instances which are collected from Yelp.com and then labeled, both need to be pre-processed. In at this point, a number of preprocessing steps are performed on the raw data obtained from web scraping or other sources to improve both its quality and its compatibility with the models chosen [25]. This includes data cleaning tasks, which entails handling missing values, removing irrelevant or redundant information, and addressing irregularities or mistakes in the information. Preprocessing is done by performing some Natural Language Processes (NLP) such as – 1. Tokenization 2. Lowercasing English letters 3. Punctuation removal 4. Stop words removal 5. Stemming. In general, the goal of data preprocessing is to raise the quality and dependability of the information, empowering more precise and productive examination and displaying processes [26].

3.3. Feature Extraction

The steps of feature extraction are enumerated in the next section. We studied in this research two different features selection methods, namely, Term Frequency (TF) and Term Frequency-Inverted Document Frequency (TF-IDF). These methods are described in the following.

3.3.1. TF

Term Frequency measures number of times a particular term t occurred in a document d. Frequency increases when the term has occurred multiple times.TF is calculated by taking ratio of frequency of term t in document td to number of terms in that particular document d [27].

3.3.2. IDF

One of the main characteristics of IDF is it weights down the term frequency while scaling up the rare ones. For example, words such as "the" and "then" often appear in the text, and if we only use TF, terms such as these will dominate the frequency count. However, using IDF scales down the impact of these terms [28].

3.3.3. TF-IDF

The Term Frequency-Inverted Document Frequency (TF-IDF) is a weighting metric often used in information retrieval and natural language processing. Pre-processing is performed before starting the

process of the TF–IDF algorithm. The following is the architecture of this statistical measure:
i) After the pre-handling step, tokenization is performed on the sentences rather than words. Then, at that point, the weight esteem is appointed to each word.
ii) Then, the calculation of word frequency is done.
iii) In this step, TF (Term Frequency) is calculated using the formula in the equation (1).

$$\mathrm{TF(w)_d} = \frac{\mathrm{n_w(d)}}{|\mathrm{d}|} \quad (1)$$

iv) After that, a table is created for the frequency of every word in every sentence.
v) Then, IDF (Inverse Document Frequency) is calculated by equation (2).

$$\mathrm{IDF(w)_d} = 1 + \log\left(\frac{|\mathrm{D}|}{|\{\mathrm{d} : \mathrm{D}|\mathrm{w} \in \mathrm{d}\}|}\right) \quad (2)$$

vi) TF-IDF is calculated by multiplying the value of equation (1) and equation (2).

$$\mathrm{TF.IDF} = \mathrm{TF(w)_d} \times \mathrm{IDF(w)_D} \quad (3)$$

vii) In these means, the typical score of all words is determined to find the edge esteem. At long last, those words are chosen in which relating score is higher than the edge esteem.

3.4. Classifiers models

Depending on our needs and our data in this study, we will use popular transformer models like BiLSTM, and many hybrid models to get high perform competitively in this research, such as: BiLSTM-CNN, BERT-CNN and RoBERTa-CNN.

3.4.1. Bi-LSTM

BiLSTM is a modification of the LSTM architecture for training in both positive and negative time directions. In other words, it is an improved version of the LSTM model where the input can be scanned simultaneously, both sequentially and in reverse order [29]. BiLSTM has two parallel layers spreading in two directions, forward and backward [30]. The final output of such a network will be the combination of the output of these two layers.

3.4.2. CNN

Convolutional Neural Network (CNN) was introduced by LeCun et al. [31] in 1999 to detect basic objects with substantial variability and is among the most significant approaches for deep learning. CNN is a feedforward artificial neural network for feature extraction and classification, among other applications. CNN is intended to be a feature extraction method in this study. CNNs have engaged in a variety of applications in the field of image processing, and more recently, they have also been utilized in the field of data classification.

3.4.3. BERT

BERT is an architecture for a deep neural network that can understand the contextual links between words in a sentence by being pre-trained on a significant amount of unlabelled text input [32]. BERT is designed to be bidirectional, meaning it can process both the left and proper contexts of a word, unlike previous language models that only processed one direction. In addition to that, it makes use of a transformer architecture, which is a self-attention mechanism [33].

3.4.4. RoBERTa

RoBERTa – The Robustly Optimized BERT Approach (RoBERTa) model was introduced in 2019 and is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture but was trained on a much larger corpus of data than BERT, with an extended training duration and improved training techniques. This allows RoBERTa to better capture complex relationships and patterns in natural language text, resulting in improved performance on a wide range of NLP tasks, including fake news classification [33].

3.4.5. BiLSTM-CNN

BiLSTM architecture used to obtain the hierarchical features that help us to extract intricate patterns of the time series characteristics, aiming to improve the effectiveness of the system [34-35]. CNN layers obtain features convolved relation among the initial hand-crafted features data in this system, while BiLSTMs predict sequences. In contrast to existing methods for converting time series to images, our model employs only the original raw data. A CNN's layers are generated using kernels that iteratively process two-dimensional sequences.

3.4.6. BERT-CNN

For the purpose of spam review detection, the BERT-CNN model combines BERT, a cutting-edge transformer-based architecture, with convolutional neural networks (CNNs). BERT succeeds at catching contextualized portrayals of text by utilizing bidirectional self-consideration components [36]. By consolidating BERT with CNNs, the model can actually encode the semantics of audits and catch neighborhood includes all the while. This mixture engineering empowers the BERT-CNN model to accomplish prevalent execution in recognizing false or misleading audits by utilizing both worldwide logical data and neighborhood designs in the survey text [37].

3.4.7. RoBERTa-CNN

TheRoBERTa-CNN model combines the power of the RoBERTa transformer-based architecture with convolutional neural networks (CNNs) for spam review detection tasks. RoBERTa, a variant of BERT model that employs a robust pre-training approach to learn deep contextualized representations of review text. By incorporating CNNs into the architecture, the model can further enhance its ability to capture local features and patterns in reviews content [38]. The RoBERTa-CNN model leverages the strengths of both architectures to effectively encode textual information and identify deceptive reviews with high accuracy and efficiency. The description of this model parameters and their values are given as follows.

4. EXPERIMENTAL RESULTS

In the context of Yelp's product reviews dataset, our experiment explores the application of deep learning with the purpose of identifying fraudulent reviews. This subsection focuses on assessing the performance of several deep learning models, encompassing hybrid approaches that blend the advantages of many architectures, as well as both conventional architectures and cutting-edge transformer-based models.

This subsection also presents our analysis for the BiLSTM (Bidirectional Long Short-Term Memory) and BiLSTM-CNN (Convolutional Neural Network) models for fake reviews detection. These models are carefully examined for their classification metrics, providing important information on how well they can distinguish between real and fake reviews. These models are meticulously scrutinized for their classification metrics, offering crucial insights into their effectiveness in discerning authentic reviews from fraudulent ones. By harnessing the power of sequential learning and convolutional operations, these models aim to capture intricate patterns within the textual data, thereby enhancing their ability to detect deceptive reviews.

Furthermore, our investigation extends to include the hybrid BERT-CNN and RoBERTa-CNN models, representing cutting-edge advancements in Natural Language Processing (NLP). These transformer-based models leverage pre-trained language representations to imbue a deeper understanding of semantic nuances within the review text, empowering them to excel in the task of fake review detection. Their attention mechanisms allow them to focus on the most informative parts of the text, thereby improving accuracy. In addition to these standalone architectures, we explore hybrid models that combine elements from different architectures. These hybrid models leverage the strengths of both traditional and transformer-based approaches, aiming to achieve superior performance in fake review detection tasks.

For the purpose of evaluating these algorithms, the precision, specificity, sensitivity, F1-score, and accuracy performance metrics were chosen. The confusion matrix measure has been used to calculate all of these metrics.

4.1. Performance Evaluation

The following evaluation metrics, which are expressed and detailed in the equations numbered (4) to (8), have been used to evaluate each of the hybrid models (deep learning & transformer-based) utilized in our proposed methods for the classification tasks of spam/fake and non-spam/truthful.

Precision=TP ⁄ (TP+FP) (4)

Specificity = TN ⁄ (TN+FP) (5)

Sensitivity (Recall) =TP⁄ (TP+FN) (6)

Accuracy=(TP+TN) ⁄ (TP+FP+FN+TN) (7)

F1-score=(2∗P∗R)⁄(P+R) (8)

Whereas,

- Precision: it is calculated by dividing the total number of positive class values predicted by the number of positive predictions.
- Specificity: it is commonly employed to calculate the total number of real negative samples (spam reviews) that were classified and predicted correctly.
- Sensitivity (Recall): it is calculated by taking the number of positive class values in the test data and

dividing it by the number of positive predictions.
- Accuracy: it is defined as the degree to which a quantity or expression is accurate.
- F1-score: it stands for striking a balance between recall and precision.

4.2. The results of experiment

This subsection presents the testing results of diverse deep learning models applied to the task of fake review detection using Yelp's product reviews dataset. We analyze the performance metrics of each model, including traditional architectures like BiLSTM and BiLSTM-CNN, along with state-of-the-art transformer-based models such as BERT-CNN and RoBERTa-CNN. Furthermore, hybrid approaches that blend the strengths of different architectures are examined.

Through this comprehensive evaluation, we aim to elucidate the effectiveness of these models in discerning genuine reviews from deceptive ones, thereby contributing valuable insights to fortify the trustworthiness of online review platforms. Tabel 3 summarizes the testing classification results of hybrid deep learning and transformers models for fake reviews detection based on Yelp product reviews dataset.

Table 3. Testing results of the independent transformers & hybrid models

Model Name	Precision	Recall	Specificity	F1-score	Accuracy	AUC
BiLSTM-CNN	0.9113	0.8864	0.9060	0.8987	0.900	0.900
BiLSTM	0.9110	0.8854	0.9057	0.8980	0.900	0.900
BERT-CNN	0.8298	0.8952	0.8269	0.8613	0.860	0.860
RoBERTa-CNN	0.8661	0.8377	0.8779	0.8517	0.8584	0.860

The experimental results underscore the BiLSTM-CNN model's remarkable efficacy in detecting fake reviews within the Yelp product reviews dataset. Boasting a precision of 0.9113, recall of 0.8864, and accuracy of 0.900, the BiLSTM-CNN model exhibits formidable classification metrics, attesting to its adeptness in discerning deceptive reviews.

The accompanying Figure 2 and 3, elucidate the model's performance through visual representations including the confusion matrix, performance plot, precision-recall curves and AUC. These graphical depictions offer nuanced insights into the model's ability to accurately classify reviews, highlighting its robustness and potential for enhancing the integrity of online review platforms.

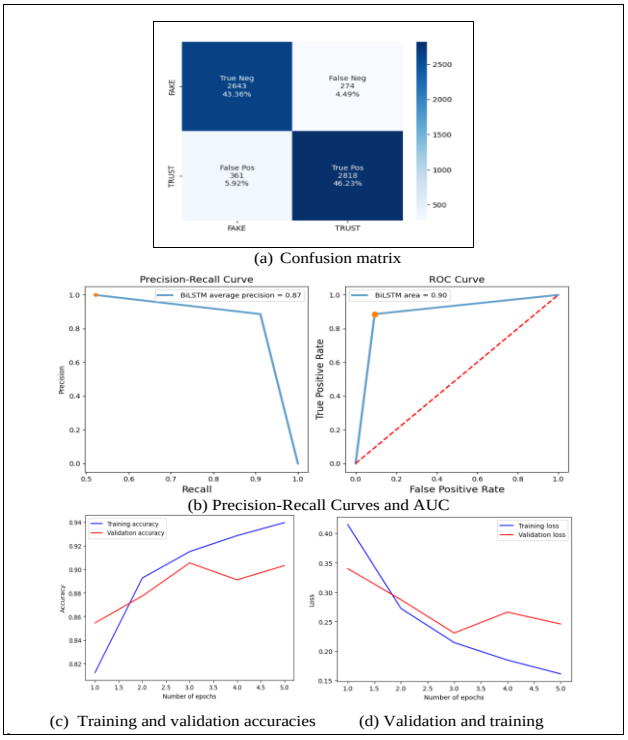

Figure 2. Confusion matrix**,** precision-recall curve and AUC, training and validation accuracies, validation and training losses of the BiLSTM-CNN model

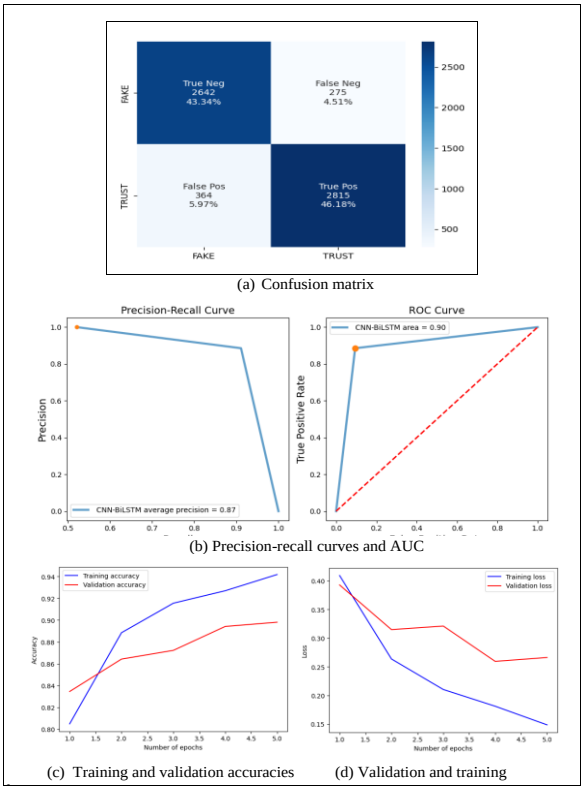

Figure 3. Confusion matrix**,** precision-recall curve and AUC, training and validation accuracies, validation and training losses of the BiLSTM model

The BiLSTM-CNN model stands out as the most effective in this context, likely due to the combination of BiLSTM's ability to capture long-range dependencies in text and CNN's strength in detecting local features. On the other hand, BERT-CNN and RoBERTa-CNN, which combine powerful pre-trained transformer-based models (BERT and RoBERTa) with CNN, also show competitive performance. Their effectiveness can be attributed to their contextual embeddings, which are known for capturing semantic meaning at a deep level, but the BiLSTM-CNN model might have an edge in handling the specific characteristics of fake reviews.

4.3. The results discussion of experiment

In the context of this experiment, which aims to detect fake reviews using deep learning models and transformer architectures applied to the Yelp product reviews dataset, the computational efficiency of these models during both training and testing phases is a crucial factor. Understanding the training and testing times provides valuable insights into the scalability and practical usability of these models in real-world scenarios. The duration of training varies among different models, with the BiLSTM model requiring approximately 264 seconds per epoch, followed by BiLSTM-CNN at around 138 to 154 seconds per epoch.

Conversely, the RoBERTa-CNN and BERT-CNN models exhibit longer training times, ranging from 218 to 285 seconds per epoch. In contrast, testing times are relatively shorter across all models, with BiLSTM, BiLSTM-CNN, RoBERTa-CNN, and BERT-CNN taking approximately 18, 11, 24, and 26 seconds, respectively, to process the test data. These disparities in training and testing times underscore the trade-offs between model complexity, computational resources, and performance metrics, emphasizing the importance of selecting the most suitable model for a given task.

Based on the provided results, it appears that the BiLSTM-CNN model exhibits the most favorable combination of training and testing times among the evaluated models. While it has achieved the highest classification performance, its efficiency in both training and testing phases makes it a compelling choice, especially in scenarios where computational resources are limited or efficiency is a priority. Therefore, considering both performance metrics and resource constraints, the BiLSTM-CNN model emerges as a strong contender for the best model in this context. Table 4 display the training and testing times for each model used in this experiment.

Table 4. Training and testing times for each model used

Model	Average Training Time(s)	Average Testing Time(s)
BiLSTM-CNN	142.8	11
BiLSTM	264	18
RoBERTa-CNN	218.8	24
BERT-CNN	230.2	26

Table 4 presents the average training and testing times for each model over five epochs. The BiLSTM model shows an average training time of 264 seconds and an average testing time of 18 seconds. Similarly, the BiLSTM-CNN model exhibits an average training time of 142.8 seconds and an average testing time of 11 seconds. Figure 4 visualizes the training and testing times for each model used in this study.

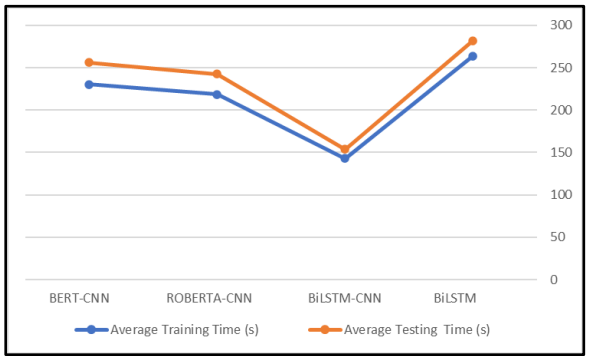

Figure 4. Visualization of training and testing times for each model

For the RoBERTa-CNN model, the average training time is 218.8 seconds, with an average testing time of 24 seconds. Lastly, the BERT-CNN model has an average training time of 230.2 seconds and an average testing time of 26 seconds.

4.4. Comparative Analysis

By conducting this comparative analysis using the same dataset and accuracy metric, we aim to provide a rigorous evaluation of our proposed models. This approach not only highlights the advancements achieved by our methods but also allows for a direct comparison with existing models. The insights gained from this analysis contribute to the broader understanding of how hybrid deep learning models can be effectively utilized for spam review detection, ultimately enhancing the reliability and credibility of online review systems. Table 5 demonstrates an comparative analysis based on the same datasets and accuracy metric.

Table 5. A comparative analysis between the results of the proposed models and existing ones usingaccuracy and same datasets

Paper Id	Dataset	Features Extraction	Feature type	Model	Results
Barbado et al. [39]	Yelp reviews dataset	TF-IDF	-Textual (Review centric features) -Behavioral (Reviewer centric features)	RF	60% 81%
Al-Adhaileh et al. [40]	Yelp reviews dataset	Word2Vec	Textual (Review centric features)	CNN BiLSTM	84% 85%
Mohawesh et al. [41]	Yelp reviews dataset	RoBERTa word embedding	Textual (Review centric features)	RoBERTa	70.2%
Our Study	Yelp reviews dataset	TF-IDF	Textual (Review centric features)	CNN-BiLSTM BiLSTM	**90%** **90%**

Table 5 cited above presents a comprehensive comparative analysis of the results obtained by our proposed models and those from existing studies, using the same datasets and accuracy as the evaluation

metric. This table includes references to various studies, the datasets used for evaluation Yelp reviews dataset, the feature extraction methods employed (such as TF-IDF, Word2Vec, and various word embedding techniques), the nature of the features (textual or behavioral), the transformer and deep learning models applied, and the accuracy achieved by each model.
CNN-BiLSTM model achieved 90% accuracy on the Yelp reviews dataset, and the BiLSTM model also achieved 90% accuracy on the same reviews dataset. Moreover, based on the provided results, the BiLSTM-CNN model exhibits the most favorable combination of training and testing times among the evaluated models. In summary, our study integrates advanced feature extraction techniques and hybrid model architectures, leading to significant improvements in accuracy for spam review detection compared to existing studies. The use of dual datasets for evaluation further strengthens the validity and applicability of our findings across different types of review data.

5. CONCLUSION

This examination incorporates the testing results and their conversation of various trials completed for counterfeit surveys location utilizing profound learning and cross breed transformer-put together models with respect to web based business spaces item audits datasets gathered from Yelp. Using Yelp datasets, it presents the confusion matrix for each model as well as the outcomes of evaluation metrics like accuracy, precision, recall, AUC, and F1-score. Additionally, a comparison of each deep learning and transformer model's learning time is provided. It contains performance plots and diagrams derived from each experiment.Overall, the results show that transformer architectures have the potential to solve the problem of online platforms detecting fake reviews. By utilizing progressed procedures for handling literary information, transformer-based models offer a promising outcomes for working on the dependability and unwavering quality of client produced content in web-based networks.

6. ACKNOWLEDGMENTS

The experimental work described in this research was conducted in the lab of Department of Computer Science, Rollwala Computer Centre, Gujarat University.

REFERENCES

[1] M. J. Zhong, L. Tan, X. L. Qu, “Identification of Opinion Spammers using Reviewer Reputation and Clustering Analysis”, International Journal of Computers Communications & Control, 14(6), 759-772, 2019.
[2] S. Yadav, D. G. Dharmela and K. Mistry, “Fake Review Detection Using Machine Learning Techniques”, JETIR, Volume 8, Issue 4, 2021.
[3] N. Jindal and B. Liu,''Opinion spam and analysis,'' in Proc. Int. Conf. Web Search Web Data Mining (WSDM), pp. 219-230, 2008.
[4] K. Goswami, Y. Park, and C. Song, “Impact of reviewer social interaction on online consumer review fraud detection,” Journal of Big Data, vol. 4, no. 1, May 2017, https://doi.org/10.1186/s40537-017-0075-6.
[5] R. Barbado, O. Araque, and C. A. Iglesias, “A framework for fake review detection in online consumer electronics retailers,” Information Processing & Management, vol. 56, no. 4, pp. 1234–1244, 2019.
[6] B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, “Towards detecting anomalous user behavior in online social networks”, In USENIX, 2014.
[7] M. Luca and G. Zervas, “Fake it till you make it: Reputation, competition, and Yelp review fraud,” Management Science, vol. 62, no. 12, pp. 3412–3427, 2016.
[8] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What Yelp Fake Review Filter might be doing?,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 7, no. 1, pp. 409–418, Aug. 2021, https://doi.org/10.1609/icwsm.v7i1.14389.
[9] S., KC & A. Mukherjee, “On the temporal dynamics of opinion spamming: Case studies on Yelp,” In Proceedings of the 25th International Conference on World Wide Web Steering Committee (pp. 369-379), 2016.
[10] S. Shehnepoor, M. Salehi, R. Farahbakhsh& N. Crespi, “NetSpam: a Network-Based spam detection framework for reviews in online social media.” IEEE Transactions on Information Forensics and Security, 12(7), 1585–1595, 2017, https://doi.org/10.1109/tifs.2017.2675361
[11] L. He, X. Wang, H. Chen, and G. Xu, “Online Spam Review Detection: A Survey of Literature,” Human-Centric Intelligent Systems, vol. 2, no. 1–2, pp. 14–30, May 2022, https://doi.org/10.1007/s44230-022-00001-3
[12] S. N. Alsubari, S. N. Deshmukh, M. H. Al-Adhaileh, F. W. Alsaade, and T. H. H. Aldhyani, “Development of integrated neural network model for identification of fake reviews in E-Commerce using

multidomain datasets," Applied Bionics and Biomechanics, vol. 2021, pp. 1–11, Apr. 2021, https://doi.org/10.1155/2021/5522574

[13] P. Gupta, S. Gandhi, and B. R. Chakravarthi, "Leveraging Transfer learning techniques- BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection," Forum for Information Retrieval Evaluation, Dec. 2021, https://doi.org/10.1145/3503162.3503169

[14] J. C. Rodrigues, J. T. Rodrigues, V. L. K. Gonsalves, A. U. Naik, P. Shetgaonkar& S. Aswale, "Machine & deep learning techniques for detection of fake reviews: A survey." In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE (pp. 1-8). February, 2020.

[15] A. Chinnalagu and A. K. Durairaj, "Comparative Analysis of BERT-base Transformers and Deep Learning Sentiment Prediction Models," in 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), 2022, pp. 874-879: IEEE.

[16] A. Alsmadiv, S. AlZu'bi, M. Al-Ayyoub & Y. Jararweh, "Predicting Helpfulness of Online Reviews."arXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.10129

[17] L. Li, W. Ren, B. Qin, and T. Liu, ''Learning document representation for deceptive opinion spam detection,'' in Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Nanjing, China: Springer, 2015, pp. 393_404.

[18] S. Zhao, Z. Xu, L. Liu, and M. Guo, ''Towards accurate deceptive opinion spam detection based on word order-preserving CNN,'' 2017, arXiv:1711.09181. http://arxiv.org/abs/1711.09181

[19] X. Wang, K. Liu, and J. Zhao, ''Handling coldstart problem in review spam detection by jointly embedding texts and behaviors,'' in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2017, pp. 366_376.

[20] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, ''DRI-RCNN: An approach to deceptive review identi_cation using recurrent convolutional neural network,'' Inf. Process. Manage., vol. 54, no. 4, pp. 576_592, 2018.

[21] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao, ''An inferable representation learning for fraud review detection with the cold-start problem,'' Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2019, pp. 1_8.

[22] C. Yuan, W. Zhou, Q. Ma, S. Lv, J. Han, and S. Hu, ''Learning review representations from the user and product-level information for spam detection,'' in Proc. IEEE Int. Conf. Data Mining (ICDM), Nov. 2019, pp. 1444_1449.

[23] N. Cao, S. Ji, D. K. W. Chiu, M. He, and X. Sun, ''A deceptive review detection framework: Combination of coarse and _ne-grained features,'' Expert Syst. Appl., vol. 156, Oct. 2020, Art. No. 113465.

[24] G. Alexandridis, I. Varlamis, K. Korovesis, G. Caridakis& P. Tsantilas. "A Survey on Sentiment Analysis and Opinion Mining in Greek Social Media."Information, 12(8), 331. 2021. https://doi.org/10.3390/info12080331

[25] T. Duyu, Q. Bing, and L. Ting, "Document modeling with gated recurrent neural network for sentiment classification". Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1422–1432, Lisbon, Portugal, 17-21 September, 2015.

[26] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah & S. B. Hassan. "Spam Review Detection Using Deep Learning. 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)". Vancouver, BC, Canada, 2019, pp. 0027-0033, https://doi.org/10.1109/iemcon.2019.8936148

[27] V. Suhasini, N. Vimala. "A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data."Turkish Journal of Computer and Mathematics Education. Vol.12 No.06, pages 5710-5723, 2021. https://doi.org/10.17762/turcomat.v12i6.10885

[28] H. Ahmed, I. Traore, & S. Saad. "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques." In Lecture notes in computer science (pp. 127–138), 2017. https://doi.org/10.1007/978-3-319-69155-8_9

[29] M. Schuster, K. K. Paliwal. "Bidirectional recurrent neural networks," IEEE Trans. Signal Process. 45, 2673–2681, 1997.

[30] Y. Yao, Z. Huang. "Bi-directional LSTM recurrent neural network for Chinese word segmentation." Int. Conf. Neural Inf. Process, pp. 345–353, 2016.

[31] A. Ghourabi, M.A. Mahmood & Q. M. Alzubi. "A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages." Future Internet, 12(9), 156, 2020. https://doi.org/10.3390/fi12090156

[32] J. Devlin, M. W. Chang, K. Lee & K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171-4186, 2018.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A robustly optimized BERT pretraining approach." 2019, arXiv:1907.11692. http://arxiv.org/abs/1907.11692

[34] M. Mendez, M. G. Merayo, M. Nunez. "Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model." Eng Appl ArtifIntell. May 2023, 121:106041. https://doi.org/10.1016/j.engappai.2023.106041

[35] N. Sharma, M. Mangla, S. Yadav, N. Goyal, A. Singh and S. Verma. "A sequential ensemble model for photovoltaic power forecasting." ComputElectr Eng. 2021, 96:107484. https://doi.org/10.1016/j.compeleceng.2021.107484

[36] A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text," Computers, Materials & Continua/Computers, Materials & Continua (Print), vol. 71, no. 2, pp. 2943–2961, Jan. 2022. https://doi.org/10.32604/cmc.2022.021671

[37] B. Zhang, "A BERT-CNN based approach on movie review sentiment analysis," SHS Web of Conferences, vol. 163, p. 04007, Jan. 2023. https://doi.org/10.1051/shsconf/202316304007

[38] K. L. Tan, C. P. Lee & K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," Applied Sciences, 13(6), 3915, 2023. https://doi.org/10.3390/app13063915

[39] R. Barbado, O. Araque, & C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," Information Processing & Management, 56(4), 1234–1244, 2019. https://doi.org/10.1016/j.ipm.2019.03.002

[40] M. H. Al-Adhaileh& F. W. Alsaade, "Detecting and analysing fake opinions using artificial intelligence algorithms," Intelligent Automation and Soft Computing, 32(1), 643-655, 2022. https://doi.org/10.32604/iasc.2022.021225

[41] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, & S. Maqsood, "Fake Reviews Detection: A survey," IEEE Access, 9, 65771–65802, 2021a. https://doi.org/10.1109/access.2021.3075573