Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.029953
Article

Tech Science Press

Leveraging Readability and Sentiment in Spam Review Filtering Using Transformer Models

Sujithra Kanmani* **and Surendiran Balasubramanian**

Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal, India
*Corresponding Author: Sujithra Kanmani. Email: sujithrakanmani@gmail.com
Received: 15 March 2022; Accepted: 30 May 2022

Abstract: Online reviews significantly influence decision-making in many aspects of society. The integrity of internet evaluations is crucial for both consumers and vendors. This concern necessitates the development of effective fake review detection techniques. The goal of this study is to identify fraudulent text reviews. A comparison is made on shill reviews *vs.* genuine reviews over sentiment and readability features using semi-supervised language processing methods with a labeled and balanced Deceptive Opinion dataset. We analyze textual features accessible in internet reviews by merging sentiment mining approaches with readability. Overall, the research improves fake review screening by using various transformer models such as Bidirectional Encoder Representation from Transformers (BERT), Robustly Optimized BERT (Roberta), XLNET (Transformer-XL) and XLM-Roberta (Cross-lingual Language model–Roberta). This proposed research extracts and classifies features from product reviews to increase the effectiveness of review filtering. As evidenced by the investigation, the application of transformer models improves the performance of spam review filtering when related to existing machine learning and deep learning models.

Keywords: Fraudulent; sentiment; readability; BERT; XLNET; roberta; XLM-roberta

1 Introduction

People's significant way of expressing themselves is now through the use of websites. Using e-commerce sites, forums and blogs, people can readily exchange their opinions on items and services. Most customers examine product and service reviews before purchasing them. Everyone on the internet increasingly recognizes the value of these online evaluations for other consumers and suppliers combined. Vendors can even build extra marketing tactics [1]. Customers found it difficult to distinguish a slanted review from an honest review written by a zealous consumer simply by looking at the review's rating because all skewed evaluations were released by unknown entities or tricksters who took a customer's name. Furthermore, [2,3] rely solely on numerical ratings to identify the presence of online review manipulation, neglecting the rich linguistic content of online reviews. We evaluate the language content of reviews and use a readability approach to detect items with altered evaluations in this research, which

goes beyond rating analysis. Although distinguishing between false and authentic reviews is complex, the writing style typically reveals specific insights that help us to discern between the both.

The problem could not be solved by physically assessing the linguistic content of a single review. Some features of the modified review were similar to another [4] to identify between honest and faked reviews. It was nearly complicated for unwitting customers to identify the manipulation of product ratings expressed in a review. Individual consumer reviews frequently reflect a personal perspective on their product experience. As a result, their writing styles should be distinct. Such distinctions represent the diversity of their culture, education, career, etc. Fake reviews are frequently more dramatic or exaggerated, whereas honest evaluations are straightforward and well-written [5]. Shill reviews are less readable than honest reviews [6,7].

Fake reviews elicit more positive or negative feelings than authentic reviews [8]. Thus sentiment analysis is also necessary in the case of feature selection for filtering the fake reviews. A review often receives several comments. We expect most of the comments to have similar feelings for genuine, authentic reviews. However, in the case of spam reviews, most consumers will likely reject the reviewer's point of view. As a result, the majority of the comments may express opposing opinions in the form of reviews.

As the influence of false reviews rises, recognizing them has become a significant issue, and ongoing study is required to handle this concerning situation. Researchers have suggested the holistic model [9,10], supervised and unsupervised machine-learning approaches [10,11] and deep-learning techniques [12,13] in recent years for identifying bogus reviews. Our study focused mostly on review text features such as readability and sentiment features. Both may be modified to produce the desired outcome. As a result, we've developed a reliable method for detecting spam reviews. We present an exposure approach that includes bag of words (BOW) test analysis, tokenization, padding and transformers models such as BERT, Roberta, XLNET and XLM_Roberta. The transformer models are evaluated using a single labeled and balanced dataset (Deceptive Opinion) where XLM_Roberta outperforms with 96% accuracy. Besides, we also tested deep-learning models mainly the Bidirectional Long Short Term Memory (Bi_LSTM) model, Convolutional Neural Network (CNN)-2D + Bi_LSTM model, CNN + LSTM (Long Short Term Memory) model, CNN + GRU (Gated Recurrent Unit) model. The CNN + LSTM model outperformed all others with the highest accuracy, 84% and minimum data loss. Our study tested machine learning models for comparison of performance with the deep-learning models, such as Logistic regression (LR), Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Random Forest (RF) Classifier. The most important contributions of this analysis are

- We developed a novel method for extracting and categorizing features included in reviews to demonstrate that readability can also be used as an effective shill review identification component.
- We conducted a careful benchmark analysis on the problem of misinformation, utilizing multiple transformers models and compared the outcomes to the State Of The Art (SOTA) models.
- We also addressed the weaknesses in the present study by using the deceptive opinion dataset and gave future directions for enhancing spam review filtering.

The rest of the paper is organized as follows; Section 2 summarizes the related works. Section 3 contains the data, methods and techniques used in this study as well as an analysis of the recommended model's structures. Section 4 goes over the experiments and their results. Finally, Section 5 concludes the study and suggests some future research directions.

2 Related Work

This section examines previous efforts to identify fake reviews using various detection algorithms. The extant literature may be classified into three categories as given below.

2.1 Transformer Based Detection Models

When a pre-trained model is used to retain contextual information concealed in raw data, the model may better understand the meaning of a letter, word, or sentence in context. BERT leverages Masking Modelling Language as a ground-breaking language model, enabling self-supervised training on massive text datasets [14]. BERT has shown SOTA performance in a diversity of NLP applications [15], with toxic comment identification [16]. Liu et al. [17] demonstrate that BERT is under-tuned and provide Roberta, an improved version of BERT, by carefully selecting the training parameters and pushing the SOTA on multiple tasks. XLM [18] improves Roberta by including Time Lapse Monitoring (TLM) in the pre-training. The XLM authors now propose XLM-R, a pre-trained model trained on big datasets in 100 languages. XLM-R achieves SOTA performance in cross-lingual detection, sequence labeling, and question answering. Several recent research [19,20] have also used XLM-R for negative text analysis and got SOTA performance.

2.2 Deep Learning Detection Models

Over the last decade, LSTM models have been acknowledged as effective models that can learn from sequence data. The capacity that makes LSTM useful is its ability to grasp long-range correlations and learn quickly from sequences of varying durations. Fraudulent card transactions have also been examined using LSTM models [21]. Bi-LSTM is a sort of recurrent neural network which is built with two hidden layers that allow bidirectional processing of the data. This is the main source of contention with LSTM. In natural language processing, Bi-LSTM has shown encouraging results [22]. Several studies demonstrate that CNN + LSTM provides a much more robust model than both CNN and LSTM separately [23]. The great performance of the CNN-LSTM model is due to the amalgamation of short and long-term feature interactions. CNN + BiLSTM model provides results over extended texts [24,25]. To efficiently create both global and local textual semantics, a CNN + GRN model may also be used to classify text. The Gate Recurrent Network may be used to evaluate user browsing history in a variety of ways [26]. As a result, the model's convergence rate can be purposely sped up [27].

2.3 Machine Learning Detection Models

Traditional machine learning algorithms for example, Support Vector Machine (SVM) [28], Random Forest [29], Multinomial Naive Bayes (MNB) [30,31] and Logistic Regression (LR) [32] rely on human feature engineering and are incapable of collecting contextual data in the toxic comment. Despite the fact that deep learning models have grown in popularity, classical models have not disappeared. Several research [33,34], suggest that LR works better in low-resource settings, whereas deep learning's potential can only be completely unleashed with enough annotated training data. Furthermore, traditional feature-based approaches maintain a model's interpretability to some extent, while most deep learning models do not. For identifying fake opinions, stylometric characteristics were applied using machine learning classifiers, namely Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) and Naive Bayes. According to trials using existing hotel reviews, exhausting stylometric traits is a promising way to detect fraudulent opinions [35]. Reference [36] combines behavioral and semantic factors in order to identify bogus reviews. It is categorized using a variety of classifiers, including Logistic Regression (LR), K-Nearest Neighbor (KNN), Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM). And the LR performs admirably in all assessment metrics [37] analyses reviewer and review-centric attributes for false review identification. It made use of the Yelp ground truth dataset, which included both actual and fraudulent review collections. Random Forests have been utilized to examine both known and unexpected traits, and the results are encouraging. Thus the leveraging of the deep neural network is taken into consideration in this work and a detailed comparative analysis is being provided in this paper, along with the expansion of future direction.

3 Materials and Methods

3.1 Dataset Description

The Kaggle Deceptive Opinion dataset is used to test the proposed analysis. This dataset contains 1600 recordings with five attributes. It's a collection of 20 real and fake hotel reviews from Chicago. The descriptions of five fields are listed in Tab. 1 below

Table 1: Description of the fields present in the deceptive opinion dataset

Fields	Description
Deceptive	There are two sorts of reviews: "truthful" and "deceptive."
Hotel	It contains the hotel's name.
Polarity	It expresses the review's emotions like positive and negative
Source	It identifies the source of the review, which comes from three sources: TripAdvisor, Mturk, and the web.
Text	It includes the reviews.

3.2 Proposed Framework for Analysis

The proposed framework for analysis expands on the existing research by incorporating Deep neural network model (Transformer models) approaches with the distinct linguistic feature of readability and sentiment mining, sets to categorize reviews from untruthful domains, thereby increasing the credibility of user-generated content available online as shown in Fig. 1 below and various phases involved in the analysis are

3.2.1 Data Acquisition

There are a few databases that contain both excellent quality real reviews and deceptive ones. Inquiring about past efforts based on the references given in Section 2, we discovered that a single labeled dataset was generally employed. The labeled dataset is obtained from Ott et al. [38]. The deceptive opinion dataset, also known as the Ott dataset, is utilized in our analysis.

3.2.2 Data Preprocessing

In this study, a series of preprocessing techniques were utilized to prepare the raw review data from the deceptive opinion dataset for computational activity. They are Tokenization, Stop words removal and Lemmatization. Tokenization divides raw text into words and phrases known as tokens. Tokenization aids in determining the meaning of the text by evaluating the word sequence. Stop words are the words which lacks meaning (e.g., "a", "an"). Any human language has an abundance of stop words. We eliminate the low-level information from our text by deleting these terms, allowing us to focus on the crucial information. In this study, all data is cleansed of stop words before proceeding with the fake review identification technique. The technique of collecting together the many inflected forms of a word so that they may be studied as a single item is known as lemmatization. Lemmatization is similar to stemming in that it adds context to words. As a result, it connects words with similar meanings to a single term. Thus the raw data goes through these three preprocessing stages.

3.2.3 Feature Selection and Extraction

This part shows a key role in the analysis phase as the feature selection decides the accuracy of the classifiers involved. This examination of the literature found two key features of distinct approaches, such as readability and sentiment for fake review detection. The extraction of the feature is based on the

review dataset and the accuracy of review spam detection is dependent on the feature engineering strategy used. As a consequence, these components must be considered in tandem for the efficient deployment of the fake review detection model and enhanced accuracy [39]. We employed the usual vector space representation techniques with TF-IDF (Term Frequency-Inverse Document Frequency) weighting [40], as well as custom tokenizers of common transformer models (BERT, Roberta and XLNet tokenizer). The major features utilized for the study is discussed as follows,

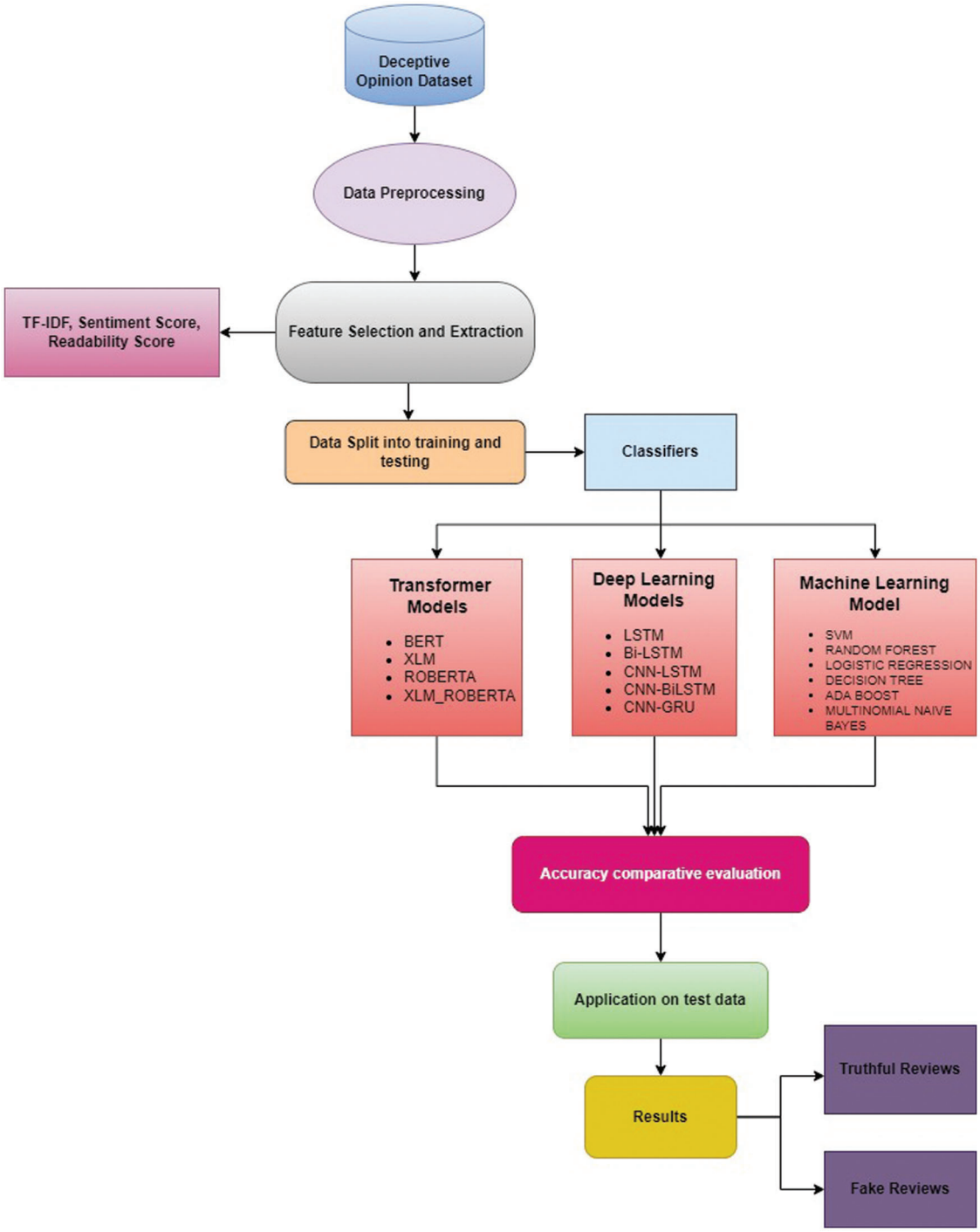

Figure 1: The proposed analysis framework

Readability Feature

In addition to the criteria listed above, we suggest an additional set of features extracted based on readability tests [41]. It is an intriguing subject of Natural Language Processing that deals with determining a document's readability. These exams determine how difficult a book is to read and comprehend. The significant cause that disturbs the authenticity of the review is the readability feature. The readability of a review's language is determined by its structural elements and captures how simple it is to interpret. The most existing system acquires low accuracy of fake reviews detection as they just use a single feature and lack the labeled experimental data. we are utilizing the key feature (i.e.,) readability as it has a hypothesis stating Shill reviews are harder to read than regular reviews [6]. We will use the following readability tests (https://readabilityformulas.com/coleman-liau-readability-formula.php) in our research dataset to analyze the importance of this feature in fake review detection and they are listed below

Flesch Reading Ease(FRE): It calculates the readability of a text on a scale of 1 to 100. When the score level goes low, the information becomes difficult to read. The mathematical formula used for calculating the readability score is shown in the Eq. (1):

$$FRE = 206.835 - (1.015 * ASL) - (84.6 * ASW) \tag{1}$$

where RE stands for Readability Ease, ASL stands for "Average Sentence Length" (i.e., the number of words divided by the number of sentences), ASW stands for Average Syllable Weighted (i.e., the number of syllables divided by the number of words). Tab. 2 shows the various score levels and their meanings.

Table 2: Flesch reading ease scores and interpretation

Score	School level	Notes
90–100	Grade 5	Very easy
80–90	Grade 6	Easy
70–80	Grade 7	Fairly easy
60–70	Grade 8	Plain english
50–60	Grade 10–12	Fairly difficult
30–50	College	Difficult
0–30	College graduate	Very difficult

The Coleman–Liau Index: This formula assesses the reading level of a text. It uses phrases and letters as variables. According to Coleman, "Letter length is a stronger predictor of readability than word length in syllables." This readability score is calculated using the mathematical procedure described in Eq. (2):

$$CLI = 0.0588L - \ 0.296S - 15.8 \tag{2}$$

where L denotes the average number of letters per 100 words and S the average number of sentences per 100 words, respectively.

SMOG: It is an abbreviation for 'Simple Measure of Gobbledygook. It is a foundation for readability. It calculates how many years of schooling the typical individual needs to comprehend a text. It works well for texts of at least 30 sentences. This was the length of text sampled during the formula's development. SMOG determines how many years of schooling an average person needs to grasp any piece of writing. This is referred to as the SMOG Grade. McLaughlin proposed calculating this using a piece of 30 phrases or more and completing the following: There are a total of 30 sentences if you count the 10 sentences at the

beginning, 10 in the middle and 10 at the end. Every word with three syllables or more is counted. Taking the square root of the integer and rounding it to the nearest ten adding three to this number. The mathematical formula used for calculating the SMOG readability score is shown in the Eq. (3):

$$SMOG\ grade = 3 + Square\ Root\ of\ Polysyllable\ Count \tag{3}$$

Dale Chall: This formula assesses word difficulty using a count of 'hard' phrases. It uses the length of the sentence and the number of 'difficult' phrases to identify a text sample's US grade level. The mathematical formula used for calculating the Dale chall readability raw score is as shown in the Eq. (4):

$$R\_score = 0.1579 * (P) + 0.0496 * L \tag{4}$$

where R_Score = Raw Score, R_Score is the reading grade of a reader who can comprehend your text. P stands for Difficult Words Percentage and L stands for Average Sentence Length in words. If P is more than 5%, then Adjusted Score will be computed as, Adjusted Score = R_Score + 3.6365; otherwise Adjusted Score = R_Score. With the adjusted score value the grade level of the reader is decided. These readability tests were conducted on the deceptive opinion dataset and the results are shown in Fig. 2. Fig. 2 contains the Average values of Truthful and deceptive reviews, where the values of the Deceptive reviews fall under the category of less readable, having a little difficulty in reading factor compared to the truthful reviews. For example, considering the Flesch reading test, the truthful reviews lie in the range of 70–80, which is fairly easy to read, but the deceptive reviews lie in the category of 60–70, which will contain Plain English making it less readable.

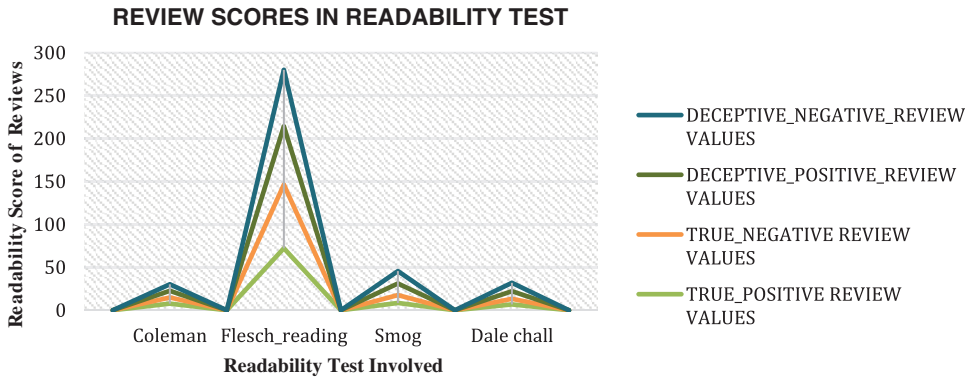

Figure 2: Reviews score in readability test

Thus by exploiting these readability tests over the deceptive opinion dataset, we can able to analyze that the fake reviews have more complexity of readability than compared to the truthful reviews, as depicted in the above Fig. 2.

Sentiment Feature

The second feature which is considered for the study is the sentiment feature, as the sentiment plays a major role in terms of classification and the VADER (Valence Aware Dictionary for Sentiment Reasoning) does the extraction of the sentiment from the review. The VADER library makes use of the polarity feature, which categorizes sentiment as positive, negative, or neutral. The compound score is calculated by summing the valence ratings of each word in the lexicon, then normalizing between extreme negative and positive. The compound score is computed using Eq. (5):

$$cs = \frac{s}{\sqrt{s^2 + \beta}} \tag{5}$$

where CS represents the computed compound score, s is the sum of all word polarity scores and β is the default value of 15. Normalization function is used such that using this hyper parameter β the maximum expected value is approximated. As stated in Eqs. (6)–(8) below, the following compound range criteria are used to classify positive, negative and neutral moods.

$$cs \geq 0.05 \quad for \quad sentiment = positive \tag{6}$$

$$cs \leq -0.05 \quad for \quad sentiment = negative \tag{7}$$

$$(cs > -0.05) \& (cs < -0.05) \quad for \quad sentiment = neutral \tag{8}$$

Reviewing evaluations to determine if they are positive, negative, or neutral. It entails predicting whether the reviews will be decent or negative based on the text's words, emoji's and review scores, among other factors. Fake reviews, according to comparable research [6], evoke more favorable or negative responses than genuine evaluations. This is because bogus evaluations are used to sway people's opinions and it's more vital to communicate ideas than it is to just give facts.

3.2.4 Transformer Models

In recent decades, transformer models have shown greater classification performance. As it employs a pre-trained model for training, the computational time is decreased and since pre-trained models are widely available as open-source, the cost of environmental setup is also lowered. This section addresses the numerous transformer models used in this investigation, which are listed below.

Bidirectional Encoder Representation from Transformers (BERT)

BERT is a deep learning language processing model with sophisticated features. By a large margin, BERT beats all previous language models. In all tiers, it operates on the collaborative left and right context phenomena. BERT is a basic yet effective tool. It shows promise in a variety of machine learning tasks. For each new model to execute a range of functions, a fine-tuned BERT model just has to add one additional layer. A veiled language model is used. MLM (Masked Language Model) is based on the phenomenon of masking random words from input and then predicting the ID of those words based on their context. MLM employs both left and right contexts, allowing for bidirectional model training. In contrast to previous language models, BERT can learn the contextual representation from both ends of the sentence. For tokenization, BERT used a 30 K vocabulary of character level Byte-Pair Encoding. The input sequence is used to produce tokens and a positional embedding. [CLS] and [SEP], two unique tokens, are added to the beginning and end of a sequence, respectively. Text categorization techniques such as Next Sentence Prediction use the [CLS] token. A separator is provided by the [SEP] token. As a result, we employed BERTBASE in our work. It isn't suitable for tasks involving ambiguous data mining or text mining. It was employed in the identification of bogus news in reference [42]. BERT was used to do sentiment analysis in reference [43] and it functioned admirably.

RoBERTa Model

The Transformers' Bidirectional Encoder Representation is abbreviated into the RoBERTa model [44]. The Transformers family, which includes the BERT and RoBERTa, was intended to address the long-range dependencies problem in sequence-to-sequence modeling. With a bigger vocabulary set of 50 K sub-word units, RoBERTa used byte-level Byte-Pair Encoding. Aside from that, the RoBERTa model improves on the BERT model by training on more data and longer sequences. The RoBERTa tokenizer includes various unique tokens, such as tokens, which denote the beginning and end of a sentence. The token is used to pad the text to achieve the maximum length of the word vector. The RoBERTa tokenizer encodes the raw text with input ids and an attention mask. The input ids represent the token indices and numerical

representation of the token. On the other hand, the attention mask is used to group the sequence together as an optional input. The attention mask indicates which tokens should be looked at and which should not.

The goal of the RoBERTa base layers is to offer a meaningful word embedding as the feature representation so that succeeding layers may readily extract useful information from it.

XLNet Model

XLNet is a BERT-based autoregressive language model that overcomes the problem of concurrently generated forecasts using BERT [45]. BERT learns by anticipating disguised words at the same time. The relationships between these predictions are not learned by predicting words simultaneously. XLNet overcame this by incorporating a permutational language model while retaining BERT's bi-directionality. It learns to anticipate words by attempting every variation of the words in a sequence. Thus, XLNet learns in a random sequence, yet in a sequential and autoregressive manner. As a result, it consistently outperforms BERT on the GLUE benchmark by 2–13 percent. Similar tokens [CLS] and [SEP] are used for classification and separation in XLNet.

XLM-RoBERTa Model

The transformer-based multilingual masked language model XLM-RoBERTa has been pre-trained on text in 100 languages and delivers cutting-edge performance in cross-lingual classification, sequence labeling and question answering [46]. XLM-RoBERTa improves on BERT by training on a larger dataset, dynamically masking tokens instead of static masking by combining a well-known preprocessing technique (Byte-Pair-Encoding) and a dual-language training mechanism with BERT to learn better relationships between words in different languages. Thus, the transformer models were utilized to boost the accuracy of spam review filtering. It is evident from the experimental results that the usage of transformer models along with readability and sentiment features gives a better future direction towards achieving the credibility of the user-generated content like reviews.

4 Experimental Analysis and Results

This section describes the experiment and the outcomes of several machine learning, deep learning, and transformer models. The tables and graphs are supplied to allow for a comparison of the models' performance.

4.1 Experimental Setup

The experiments are written in Python 3.6.9 in Google Colab to make use of the GPU's computing capabilities. Numpy 1.18.5 and Huggingface 3.5.1 are used for data preparation and tokenization. Huggingface 3.5.1 is also used to implement the pre-trained transformers. Scikit-learn 0.23.2 is used to implement the Machine learning model. Pytorch 1.7.0 or Tensorflow 2.3.0 are used to create deep learning models. Matplotlib 3.2.2 is used to create the graphs.

4.2 Experimental Evaluation of Machine Learning Models

The reviews are categorized as fake or non-fake using several Machine learning classifiers and assessed for accuracy. Logistic regression classifiers excel in accuracy, whereas Support Vector Machine and Multinomial Naive Bayes outperform the other classifiers. The dataset is alienated into multiple train and test sections, and the accuracy is reported as given in Tab. 3. The results of various classifiers involved in the fake and truthful reviews and their performance score are being tabulated as shown in Tab. 4.

Table 3: Accuracy evaluation details for classifiers involving readability and sentiment features

Classifiers involved	Training and testing ratio (%) for classification accuracy			
	60:40 ratio (%)	70:30 ratio (%)	80:20 ratio (%)	90:10 ratio (%)
Support vector machine	85.9	83.9	86.2	86.2
Random forest	78.4	74.5	74	78
Decision tree	65.3	67.5	64	68
Logistic regression	86.5	85.6	87	87.7
Ada-boost	79.2	80	78	80
Multinomial naïve bayes	81.8	83.3	83.1	85

Table 4: Performance score of classifiers

Classifiers	Truthful reviews			Fake reviews		
	Precision (%)	Recall (%)	F-score (%)	Precision (%)	Recall (%)	F-score (%)
Support vector machine	85	81	83	82	86	84
Random forest	73	81	77	79	71	75
Decision tree	67	67	67	67	67	67
Logistic regression	86	84	85	84	86	85
Ada-boost	71	85	81	83	75	79
Multinomial naïve bayes	91	76	83	80	93	86

Thus this paper studied various machine learning classifiers for classifying the fake and truthful reviews. Logistic regression excels in accuracy among the classifiers. Multinomial Naïve Bayes (MNB) and Support Vector Machine performed better than other classifiers.

4.3 Experimental Evaluation of Deep Learning Models

The Deceptive opinion dataset was used in the experiment. Compared to traditional models like LSTM, BI-LSTM, CNN + Bi-LSTM and CNN + GRU, the proposed combinational recommended model CNN and LSTM with sentiment intensity value offer superior results. Furthermore, the findings of this hybrid technique surpass the sentiment intensity values-based model. The proposed hybrid (CNN-LSTM) model outperforms existing methods in terms of accuracy. The loss function of the recommended model outperformed other models in terms of performance measures. Figs. 3–7 [47] show the correlation accuracy estimations of multiple models for the misleading opinion dataset and the results are summarized in Tab. 5.

Fig. 3 depicts the LSTM model with an accuracy value of 80.5 percent. Fig. 4 depicts the Bi-LSTM model with an accuracy of 82.5 percent. The CNN-BiLSTM model has a 59 percent accuracy, as seen in Fig. 5. On the deceptive opinion dataset, this combination has the lowest accuracy percentage. Fig. 6 depicts the CNN-GRU model with a 66 percent accuracy. The suggested CNN-LSTM combinational model is presented in Fig. 7 and it outperforms the current models in terms of accuracy by 87.7 percent involving readability and sentiment features. The first fifty review sentences from the testing set are

utilized as input and the deception of these phrases is predicted using the model we provide. The model's accuracy is then compared against the accuracy of several inspired deep learning models. The anticipated percentages of accuracy and deception are being computed.

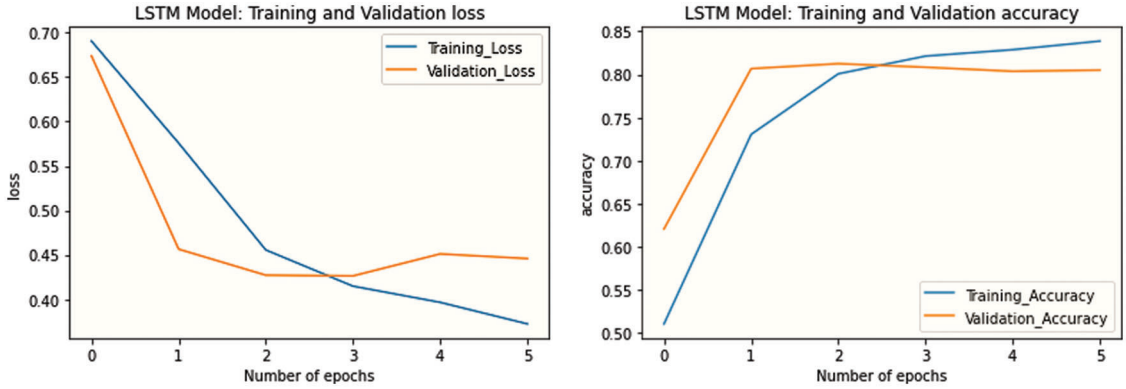

Figure 3: LSTM model with loss and accuracy values

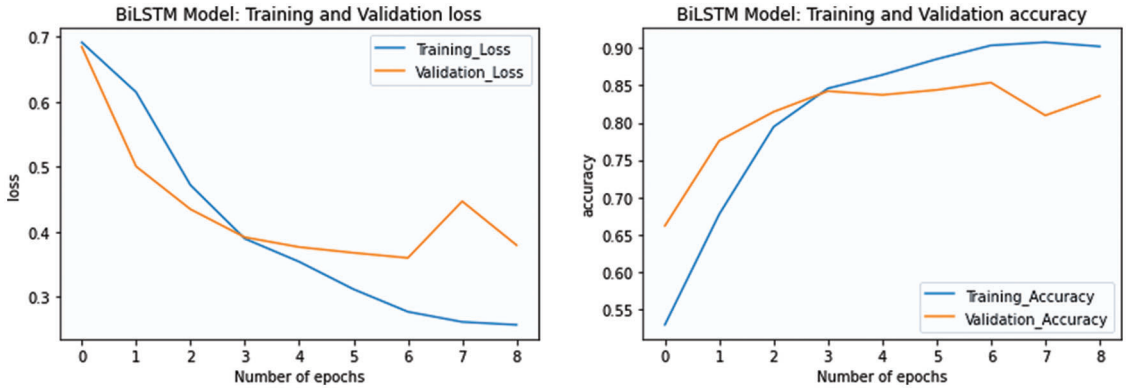

Figure 4: Bi-LSTM model with loss and accuracy values

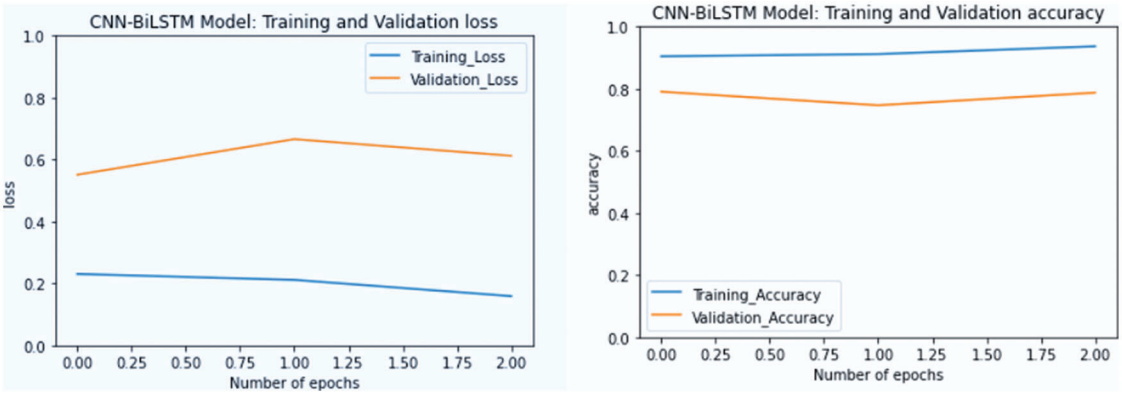

Figure 5: CNN + Bi-LSTM model with loss and accuracy values

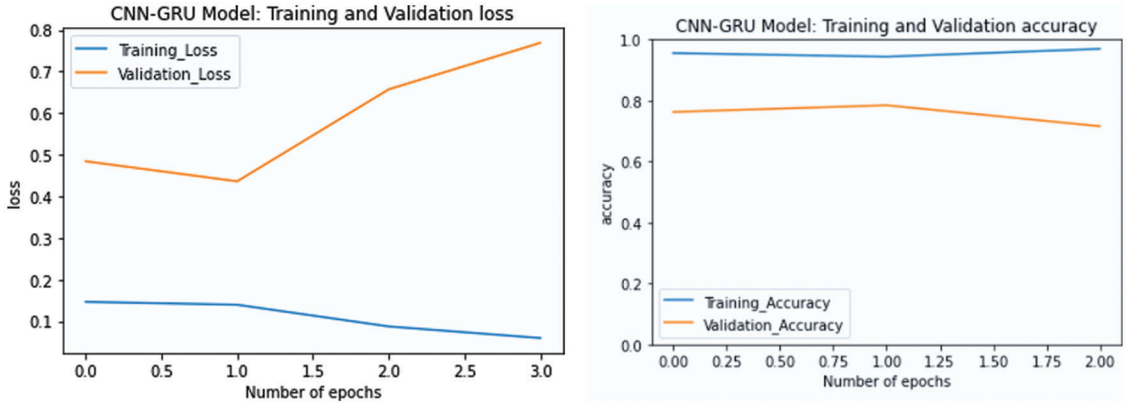

Figure 6: CNN + GRU model with loss and accuracy values

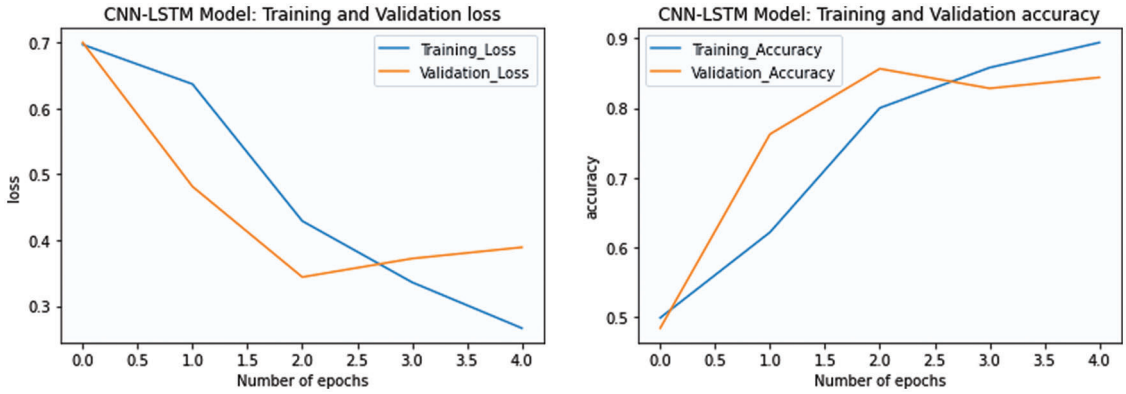

Figure 7: CNN + LSTM model with accuracy and loss curves

Table 5: Accuracy evaluation details for deep learning models involving readability and sentiment features

Deep learning model	Accuracy percentage involving sentiment features only [47]	Accuracy percentage involving readability and sentiment features
LSTM	80.5	80.5
Bi-LSTM	82.5	82.5
CNN + Bi-LSTM	49	59
CNN + GRU	42	66
CNN + LSTM	83.7	87.7

This article investigated the use of deep learning models, finding that a hybrid mix of CNN and LSTM with readability and sentiment features outperforms other deep learning models such as LSTM, Bi-LSTM, CNN + Bi-LSTM and CNN + GRU in terms of accuracy.

4.4 Experimental Evaluation of Transformer Models

The study aims to exploit the transformer models over the deceptive opinion dataset. A range of learning rates between $1e-3$ and $5e-5$ will be examined, along with batch sizes ranging from 16 to 32. The models will be trained using Adam optimization and cross-entropy as a loss function. The following settings will be

fine-tuned: Number of batches: [16,32] and Rate of learning: [1e − 3, 5e − 3, 1e − 4, 5e − 4, 1e − 5, 5e − 5]. Below Tab. 6 shows the results of transformer models on the deception dataset. Again the models were evaluated on the accuracy, recall, precision and f1-score as it is given in Fig. 8. All models were run on five different train-test splits.

Table 6: Performance details for transformer models

Models	Accuracy	Precision	Recall	F-score	Epoch
BERT	91.2	93	89	91	5
XLNET	94.3	94.7	94	94.3	5
RoBERTa	97.13	98	96.4	97	5
XLM-RoBERTa	98.2	97.8	98.6	98.2	5

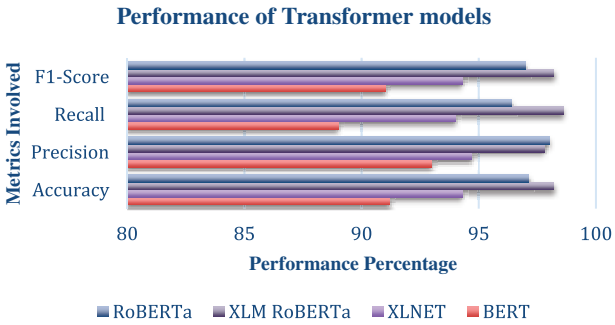

Figure 8: Performance of transformer models

4.5 Findings and Contribution

The study's goal was to investigate how we can employ pre-trained transformers to detect spam reviews. To begin, we experimentally investigated the best combination of machine learning and deep learning models. Next, we demonstrated that combining transformer-based classifiers improves performance against spam review filtering. BERT, RoBERTa, XLNet pre-trained language models were used. RoBERTa and XLNet were able to classify false reviews more effectively. Overall, RoBERTa-based combination models outperformed all others. In machine learning, logistic regression gave better excellence, and in the case of deep learning, the CNN-LSTM combination outperformed the other models.

5 Conclusion and Future Direction

This study looked into how different pre-trained transformers may be used to identify online spam reviews. Furthermore, this study added to current research by assembling the best models utilizing readability and sentiment features for spam review identification. In conjunction with all classification models, RoBERTa and the combination of RoBERTa with XLM outperformed BERT in detecting spam reviews. These transformers are more sophisticated and as a result, can better convey the review's content. Additionally, the transformer model outperformed the machine learning and deep learning models. Thus transformers in depth might be quite valuable for natural language processing as it saves time with the pre-trained models by achieving excellence of efficiency. The Future direction is towards working on the unavailability of the labeled dataset where the behavioral features will be taken for

considering the fake review filtering. As the reviews are of user-generated content, they may consist of multilingual categories, and thus multilingual review spam detection will be explored.

Funding Statement: The authors received no specific funding for this study

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Xue, F. Li, H. Seo and R. Pluretti, "Trust-aware review spam detection," *2015 IEEE Trustcom/BigDataSE/ISPA*, Helsinki, Finland, vol. 1, pp. 726–733, 2015.

[2] N. Hu, L. Liu and V. Sambamurthy, "Fraud detection in online consumer reviews," *Decision Support Systems*, vol. 50, no. 3, pp. 614–626, 2011.

[3] N. Hu, I. Bose, Y. Gao and L. Liu, "Manipulation in digital word-of-mouth: A reality check for book reviews," *Decision Support Systems*, vol. 50, no. 3, pp. 627–635, 2011.

[4] M. K. P. Orlow, H. A. Taylor and F. L. Brancati, "Readability standards for informed-consent forms as compared with actual readability," *The New England Journal of Medicine*, vol. 348, no. 8, pp. 721–726, 2003.

[5] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin and J. F. Nunamaker, "A comparison of classification methods for predicting deception in computer-mediated communication," *Journal of Management Information Systems*, vol. 20, no. 4, pp. 139–166, 2004.

[6] A. Vartapetiance and L. Gillam, "I don't know where he is not: Does deception research yet offer a basis for deception detectives?" in *Proc. of the Workshop on Computational Approaches to Deception Detection*, Avignon, France, pp. 5–14, 2012.

[7] T. Ong, M. Mannino and D. Gregg, "Linguistic characteristics of shill reviews," *Electronic Commerce Research and Applications*, vol. 13, no. 2, pp. 69–78, 2014.

[8] M. Chakraborty, S. Pal, R. Pramanik and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: A survey," *Information Processing & Management*, vol. 52, pp. 1053–1073, 2016.

[9] Y. Li, Z. Qin, W. Xu and J. Guo, "A holistic model of mining product aspects and associated sentiments from online review," *Multimedia Tools and Applications*, vol. 74, no. 23, pp. 10177–94, 2015.

[10] J. K. Rout, S. Singh, S. K. Jena and S. Bakshi, "Deceptive review detection using labeled and unlabeled data," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3187–211, 2017.

[11] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi supervised and supervised learning," in *2019 Int. Conf. on Electrical, Computer and Communication Engineering*, Cox's Bazar, Bangladesh, pp. 1–5, 2019.

[12] S. Girgis, E. Amer and M. Gadallah, "Deep learning algorithms for detecting fake news in online text," in *2018 13th Int. Conf. on Computer Engineering and Systems*, Cairo, Egypt, pp. 93–7, 2018.

[13] A. Hassan and A. Mahmood, "Deep learning approach for sentiment analysis of short texts," in *2017 3rd Int. Conf. on Control, Automation and Robotics*, Nagoya, Japan, pp. 705–10, 2017.

[14] G. Song, D. Huang and Z. Xiao, "A study of multilingual toxic text detection approaches under imbalanced sample distribution," *Information*, vol. 12, no. 5, pp. 205, 2021.

[15] D. Jacob, C. Ming-Wei and L. Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, vol. 1, pp. 4171–4186, 2019.

[16] M. Mozafari, R. Farah bakhsh and N. Crespi, "A Bert-based transfer learning approach for hate speech detection in online social media," in *Int. Conf. on Complex Networks and Their Applications*, Springer, Switzerland, pp. 928–940, 2019.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi *et al.,* "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[18] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv:1901.07291, 2019.

[19] T. Ranasinghe, M. Zampieri and M. MUDES, "Multilingual detection of offensive spans," arXiv, arXiv:2102.09665, 2021.

[20] Roy, S. Gosh, U. Narayan, T. Raha, Z. Abid *et al.,* "Leveraging multilingual transformers for hate speech detection," arXiv:2101.03207, 2021.

[21] B. Wiese and C. Omlin, "Credit card transactions, fraud detection, and machine learning: Modelling time with lstm recurrent neural networks," in *Innovations in Neural Information Paradigms and Applications*, Berlin, Heidelberg: Springer, pp. 231–268, 2009.

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, 1997.

[23] Y. A. Amrani, M. Lazaar and K. E. E. Kadiri, "Random forest and support based hybrid on vector intelligent machine approach to sentiment analysis," *Procedia Computer Science*, vol. 127, pp. 511–520, 2018.

[24] M. Rhanoui, M. Mikram, S. Yousfi and S. Barzali, "A cnn-bilstm model for document-level sentiment analysis," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832–847, 2019.

[25] L. Zhang and F. Xiang, "Relation classification via bilstm-cnn," *Data Mining and Big Data*, no. 10, pp. 373–382, 2018.

[26] S. Okura, Y. Tagami, S. Ono and A. Tajima, "Embedding-based news recommendation for millions of users," in *Proc. of the 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, pp. 1933–1942, 2017.

[27] B. Liu, Y. Zhou and W. Sun, "Character-level text classification via convolutional neural network and gated recurrent unit," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 8, pp. 1939–1949, 2020.

[28] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Sheffield, UK, vol. 25, no. 29, pp. 468–469, 2004.

[29] I. Alfina, R. Mulia, M. I. Fanany and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 Int. Conf. on Advanced Computer Science and Information Systems (ICACSIS)*, Indonesia, pp. 233–238, 2017.

[30] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Twenty-Seventh AAAI Conf. on Artificial Intelligence*, USA, vol. 27, pp. 1621–1622, 2013.

[31] L. Chen, L. Hong and J. Liu, "Analysis and prediction of new media information dissemination of police microblog," *Journal of New Media*, vol. 2, no. 2, pp. 91–98, 2020.

[32] M. A. Saif, A. N. Medvedev, M. A. Medvedev and T. Atanasova, "Classification of online toxic comments using the logistic regression and neural networks models," in *AIP Conf. Proc.*, New York, NY, USA, vol. 2048, pp. 060011–060014, 2018.

[33] X. Huang, L. Xing, F. Dernoncourt and M. J. Paul, "Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition," arXiv preprint arXiv:2002.10361, 2020.

[34] S. S. Aluru, B. Mathew, P. Saha and A. Mukherjee, "Deep learning models for multilingual hate speech detection," arXiv:2004.06465, 2020.

[35] S. Shojaee, M. A. A. Muradt, A. B. Azman, N. M. Sharef and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features," in *2013 13th Int. Conf. on Intelligent Systems Design and Applications*, Salangor, Malaysia, pp. 53–58, 2013.

[36] X. Wang, X. Zhang, C. Jiang and H. Liu, "Identification of fake reviews using semantic and behavioral features," in *2018 4th Int. Conf. on Information Management (ICIM)*, Oxford, UK, pp. 92–97, 2018.

[37] J. Fontanarava, G. Pasi and M. Viviani, "Feature analysis for fake review detection through supervised classification." in *2017 IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA)*, Tokyo, Japan, pp. 658–666, 2017.

[38] M. Ott, Y. Choi, C. Cardie and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," arXiv preprint arXiv:1107.4557, 2011.

[39] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain and M. Kaleem, "Spam review detection techniques: A systematic literature review," *Applied Science*, vol. 9, no. 987, pp. 1–26, 2019.

[40] V. Maslej, M. Sarnovský, P. Butka and K. Machová, "Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification," *Applied Science*, vol. 10, no. 8631, pp. 1–26, 2020.

[41] M. P. O'Mahony and B. Smyth, "Using readability tests to predict helpful product reviews," in *RIAO 2010 the 9th Int. Conf. on Adaptivity, Personalization and Fusion of Heterogeneous Information*, Paris, France, pp. 164–167, 2010.

[42] R. K. Kaliyar, A. Goswami and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11765–11788, 2021.

[43] C. Sun, L. Huang and X. Qiu, "Utilizing bert for aspect based sentiment analysis via constructing auxiliary sentence," arXiv preprint arXiv:1903.09588, 2019.

[44] H. Al-Jarrah, R. Al-Hamouri and A. S. Mohammad, "The impact of roberta transformer for evaluation common sense understanding," in *Proc. of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), pp. 521–526, 2020.

[45] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov *et al.,* "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5753–5763, 2019.

[46] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek *et al.,* "Unsupervised cross-lingual representationl learning at scale," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL (online), pp. 8440–8451, 2020.

[47] R. S. Kanmani and B. Surendiran, "Boosting credibility of a recommender system using deep learning techniques-an empirical study," *International Journal of Engineering Trends and Technology*, vol. 69, no. 10, pp. 235–243, 2021.