National College of Ireland

Deep Learning Approaches for Identifying fake Reviews in E-Commerce Platforms

MSc Research Project
MSc in Data Analytics

Shiva Vasineni
Student ID: 23201274

School of Computing
National College of Ireland

Supervisor: Dr.William Clifford

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shiva Vasineni
Student ID:	23201274
Programme:	MSc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr.William Clifford
Submission Due Date:	29/01/2025
Project Title:	Deep Learning Approaches for Identifying fake Reviews in E-Commerce Platforms
Word Count:	8039
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shiva Vasineni
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning Approaches for Identifying fake Reviews in E-Commerce Platforms

Shiva Vasineni
23201274

Abstract

The accumulation of fake reviews are currently widespread among the various e-commerce platforms has created a problem of credibility and distorted purchasing choices. This research's main objective is to use efficient machine-learning and deep learning approaches to fake review identification. The aim is to classify a given review as fake or real using natural language processing (NLP) techniques and using both the conventional machine learning models along deep learning methods. The dataset contains both real and fake reviews where OLAMA model is used to generate fake reviews and combined with original reviews. To estimate the performance of the proposed model, another test set of 5,000 samples including fake and genuine reviews is used for testing. The machine learning baseline models such as Decision Tree, Random Forest, and Naive Bayes classifiers are used; on the other hand, deep learning models such as LSTM and CNN + BiLSTM are used for comparison purposes. The main goal of the project is to improve the performance of fake review detection which in turn would help improve the overall experience of users of e-commerce platforms.

1 Introduction

The exponential growth of e-commerce has greatly changed the world of retailing and introduced new opportunities for everyone. However, this has also led to various problems that arose from the fast expansion mainly in the review sites for online products. These reviews which are very popular and serve as major influencers of the consumers buying decisions have also been widely susceptible to fake reviews. Such positive or negative fakes are generally used to mislead potential buyers, change their perception of the product, and erode their confidence. Consequently, the authenticity of the user-contributed content on e-commerce sites has been questioned, which increased the general awareness of the credibility of online reviews in influencing consumers' buying decisions (Alsubari et al.; 2023).

Most fake reviews are usually from people who have a negative attitude toward a particular product, or else they are computer programs that have been designed to write positive things about a product (Zhang et al.; 2020). However, while some may be generally fake and easily spotted due to language that is overemotional or simply not coherent, some are written so professionally that they nearly mimic authentic users' feedback. This poses a major problem to e-commerce sites as they are required to offer consumers accurate information while at the same time dealing with a never-ending stream of content

from the users. The present approaches to manual detection are user reporting and simple rule-based filters, which are insufficient to deal with this problem at a larger level. Since fake reviews are getting more sophisticated, stronger, and automated approaches are necessary to handle the issue more efficiently (Bathla and Kumar; 2021; Tufail et al.; 2022).

In this regard, machine learning techniques, especially deep learning methods open a lot of potential to enhance the efficiency and effectiveness of fake review detection. LSTM and CNN networks are best suited for learning, analyzing, and making predictions on complex textual data that has higher dimensionality. RNN and LSTMs are used in the current study because they are effective in handling textual information where context and sequence are key factors (Sumathi et al.; 2021; Thuy et al.; 2024). On the other hand, the CNNs are very efficient for localized feature extraction and n-grams in text which are very useful in detecting a small character that may lead to a fake review. One of the approaches that can be followed to create new complex models that are more powerful than the LSTMs or CNNs alone is the use of more complex structures that integrate both the LSTMs and the CNNs so that the model may incorporate both the global and local contextual information and thus be able to differentiate between the real and the fake content.

The future of detecting fake reviews in e-commerce may be Large Language Models (LLMs) (Naveed et al.; 2023; Mann et al.; 2020; Achiam et al.; 2023). These models are trained with large amounts of written language data and have a developed capability of understanding the human language. It is also revealed that LLMs can be used to extract fine-grained information concerning the language structure, polarity, and contextual relations within the reviews where such factors sometimes hold the real or fake signals. Because of the long-range dependencies and stylistic differences in writing detected by LLMs, they can improve the identification of fake, including sophisticated, reviews. Moreover, LLMs can be trained afresh, specifically to be able to counter the new strategies employed in conceiving fake reviews which outweighs the usefulness of LLMs when it comes to the recurrent and incessant battle against fake content in online platforms.

The objective of this research is to develop and compare deep learning models for detecting fake reviews in e-commerce platforms. In particular, this work will compare LSTM and CNN-based structures for fake reviews' detection from a dataset containing real and artificially created fake reviews. The ultimate goal of this work is to enhance the discrimination capability of fake review detection models based on the deep learning approach to capture subtle language differences between fake and genuine reviews. By doing extensive experiments including the comparison of traditional machine learning models like Decision Trees, Random Forests, and Naive Bayes Classifier, this research aims to establish the efficacy of deep learning models in addressing the ongoing challenge of fake review detection thus enhancing online shopping credibility when consumers are making a purchase online.

This research contributes to the field of fake review detection by presenting a comparative analysis of deep learning models, particularly CNN-BiLSTM, against traditional machine learning models. Unlike prior studies that primarily rely on either classical machine learning techniques or standalone deep learning models, this work integrates convolutional networks with bidirectional long short-term memory networks to capture both local and global text dependencies. By leveraging a synthetic dataset generated using the OLAMA model alongside real-world reviews, this study enhances the robustness

of fake review classification models. Furthermore, it systematically evaluates the impact of different architectures on performance metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive benchmark for future studies.

Another key contribution of this study is the exploration of the use of large language models (LLMs) in detecting fake reviews. Given the increasing sophistication of automatically generated deceptive content, this research highlights the potential of transformer-based architectures in identifying nuanced textual patterns that distinguish real from fake reviews. By experimenting with different data preprocessing techniques, including TF-IDF vectorization and stopword removal, the study also refines the feature extraction process to optimize model performance. The findings contribute to the ongoing discourse on enhancing online review credibility, offering practical implications for e-commerce platforms, regulatory bodies, and AI practitioners focused on mitigating review fraud.

2 Literature Review on Fake Review Detection

The proliferation of fake reviews has become one of the biggest issues currently being experienced on e-commerce platforms and is a threat as far as customer confidence and their decision making is concerned. These reviews may manipulate actual evaluation of products and services; and mislead customers, thereby eroding the legitimacy of online shopping platforms (Fusilier et al.; 2015). That is why it is necessary to identify fake reviews and let consumers make the right purchasing decisions. Several approaches have been suggested in the literature, starting with conventional rule-based models and ending with modern machine learning and deep learning models to deal with the problem of identifying fraudulent reviews (Pavlinek and Podgorelec; 2017). As e-commerce emerges as the latest trending topic in today's society, this relatively new research area has endeavored to design more efficient and scalable approaches for identifying fake reviews from massive data (Alsharif; 2022).

2.1 Earlier Approaches to Fake Review Detection

In the early research on fake review detection, the major approach primarily relies on the rule-based model that is based on the use of language, sentiment analysis, and the number of keywords or phrases in the fake and real reviews (Li et al.; 2014). Such methods usually used plain text characteristics based on BoW or n-grams to search for 'abnormal' patterns (Pavlinek and Podgorelec; 2017). In addition, other topic models like Latent Dirichlet Allocation (LDA) were used to identify latent topics that existed within the given reviews which could be employed in the identification of common contradictions typical of fake reviews (Yelundur et al.; 2019). In addition, users' behavior was also taken into account in the reviewing process, and the reputation scores for evaluating the reviewers and for detecting outliers (Rayana and Akoglu; 2015). The challenges that were experienced with these approaches were that the feature had to be designed by hand, and the model could not capture the relations that were present in the new and more advanced fake review techniques. Researchers then attempted to try other techniques in the machine learning approach to cope with the fact that fake review detection had become a tougher task (Fusilier et al.; 2015).

2.2 Machine Learning-Based Approaches

At present, ML methods have been employed in detecting fake reviews because they encompass the feature learning process thereby reducing feature engineering. Such familiar models as Decision Trees, Random Forests, and Naïve Bayes classifiers are used in fake review detection while such features as the length of the review, its sentiment, and activity levels of the user are used for the categorization (Yelundur et al.; 2019; Zhang et al.; 2020). Further, although labeled data have some drawbacks, the methods for learning from PU (Positive-Unlabeled) data have been used for learning in cases where only positive samples and a set of unlabeled samples are available (Fusilier et al.; 2015). Other works have also employed the hybrid of different classifiers to enhance the reliability and accuracy of the model since different characteristics of data can depict different facts (Liu et al.; 2021). The authors have also employed methods of feature fusion that combine textual features, sentiment analysis scores as well as other metadata including the history of the reviewer (Li et al.; 2014). These models have demonstrated fairly good performance in the detection of fake reviews although they are still weak in the sense that the features they have been designed to capture are only a subset of the features that are inherent in natural language and that the features have to be predetermined (Yelundur et al.; 2019). The shift to deep learning-based methods is described to address these challenges since they are not preceded by the extraction of features from the raw data (Liu et al.; 2021).

2.3 Deep Learning-Based Approaches

A recent trend in fake review detection has been recommended to employ deep learning methods since they outperform in extracting features from the textual data without the need to determine the features manually. Especially, RNNs (Recurrent Neural Networks), particularly LSTM (Long Short-Term Memory) for fake review detection have been used more frequently because the sequential feature in the text data is useful for deception detection (Baishya et al.; 2021; Alsharif; 2022). For example, LSTM networks are useful in detecting linguistic and context features that distinguish fake reviews from real ones making it a suitable model for use here (Zhang et al.; 2023). Furthermore, there are also other techniques such as applying attention mechanisms into the deep learning models to make the system pay more attention to the important features of the text which contribute to a high accuracy rate (Thuy et al.; 2024). Furthermore, even with recent models such as BERT (Bidirectional Encoder Representations from Transformers), the performance has been increased sharply by training on a huge language model that captures high abstract meanings (Thuy et al.; 2024). These models have been very successful in identifying both local and global dependencies from the review texts, thus they are very relevant in fake review detection in e-Commerce (Alsubari et al.; 2022). However, while deep learning models yield high performance, they critically depend on large supplies of labeled data and large-scale computation, which are problematic for practical adoption. Nevertheless, these models still encounter new challenges in handling the very heterogeneous and dynamically changing environment of fake reviewing strategies.

2.4 Research Gaps

While existing machine learning and deep learning models have demonstrated strong performance in fake review detection, they often face challenges in handling the complexity and diversity of fake reviews, particularly those employing sophisticated tactics to mimic

genuine ones. Most of the existing methods the use pre existing datasets that might not be sensitive to the dynamics of fraudsters' activities. Furthermore, generally, deep learning models work well only when large, labeled datasets are available for their learning, and such sets are not always accessible in sufficient quantity. One significant contribution to addressing these challenges is the generation of fake reviews using advanced language models like GPT. By synthesizing fake reviews from such models, it is possible to create a more diverse and dynamic dataset that better reflects the variety of deceptive tactics used in online reviews. This approach can facilitate the development of more adaptive, robust models that are better equipped to detect emerging forms of manipulation in real-world applications.

3 Fake Review Generation

In this study, the OLAMA model was used for the synthesis of realistic fake reviews because of its high text quality and ability to generate human like text. Since they are trained to generate diverse andcoherent content , OLAMA can be very useful in artificially generating different reviews regarding the product with different kinds of sentiments including positive, negative, and neutral. This realism is very important in the training of fake review detection models because the model can be trained by detecting properly. Further, OLAMA being an open source has several advantages over commercial models such as GPT, which require ongoing charge for subscription. OLAMA is free to use, customizable and openness while not being limited to the number of reviews to be generated for a large number of consumers. This makes OLAMA optimal for genearting varied realistic data sets across the e-commerce product categories and allows more extensive testing, unlike the paid models.

3.1 Overview of OLAMA

OLAMA (Open Language Model for Automatic Text Generation) is a highly versatile pre-trained language model, which is intended for text generation using prompts. It is based on the deep learning approach: the transformer architecture that helps the model to process coherent and diverse text in various domains and conditions. OLAMA can generate various types of reviews, such as product reviews, articles, stories, etc., which is very suitable for the task that requires the creation of a large number of reviews. In this project, OLAMA is used to produce fake reviews for e-commerce products in a similar style to users' generated content. Since OLAMA generates prompts for each category, the reviews produced are of various sentiments – positive, neutral, and negative – which makes the generated content useful for training models to identify fake reviews.

3.2 Review Generation Process Using OLAMA Model

The fake reviews creation process starts with the identification of product categories like Kitchen, Electronics, Sports & Outdoors, where every category is given a detailed specification. These prompts tell the OLAMA model to write 2,000 reviews for each category depending on the sentiment which can be positive, negative, or neutral. The reviews are supposed to focus on the characteristic features of the product, advantages, and shortcomings, general sentiments, and impressions in brief 1-3 line prompts. Once OLAMA receives the prompts, it produces the reviews in sets, where each of the reviews

is pulled using regex. The reviews are then filtered from the response generated by the model which only captures the review enclosed in quotes.

3.3 Review Collection and Output Generation

To achieve the number of reviews per category (2,000), the review generation process is conducted in a loop until the number of reviews is sufficient. If a batch of products does not receive many reviews, more requests are sent to cover the shortage. The reviews are then cleaned for quality, where the duplicates are removed, and the reviews that are too short or those that are not coherent are discarded. The generated reviews are then put in a structured DataFrame together with the product categories of the reviews generated. Last, the DataFrame is written to an Excel file so that the dataset is handy for other analysis or even model building. This approach allows generating a large and diverse set of fake reviews at a relatively low cost, which will be necessary for evaluating and training fake review identification models.

3.4 Dataset Overview

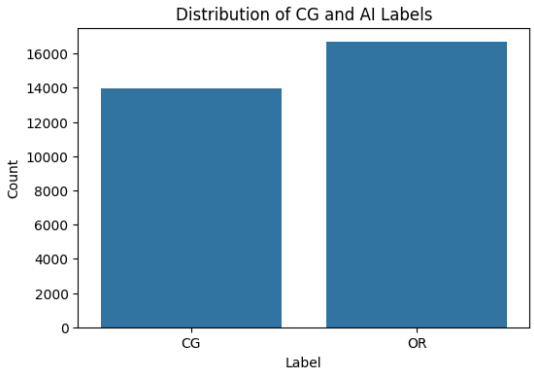
The dataset for this project is intended for training and testing machine Learning models for fake review detection. The dataset totals 30,000 samples for training and 5,000 samples for testing. The training data and testing data is split into two key classes first one is original or real review, and the second, being fake. For 30,000 training samples, the number of original reviews is 16,000, while the number of fake reviews is 14,000. By having a good ratio of fake and real reviews, the training of the models is made possible such that the models will be trained to differentiate the actual and the synthetic reviews using certain linguistic features.

3.5 Training, Validation, and Testing Split

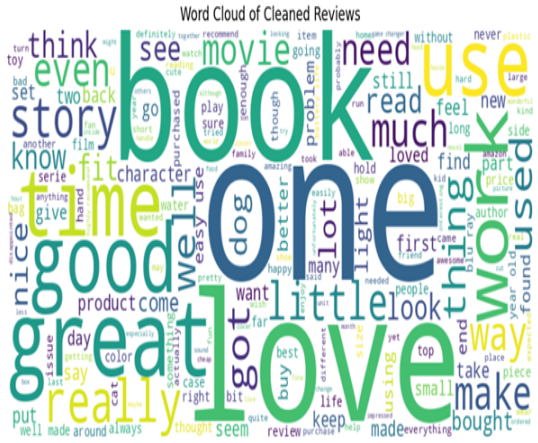
The training set is then split further to contain a validation set that is used in hyperparameters tuning and choosing the right model. In particular, the size of the validation set is 15% of the size of the training data, which is necessary for tuning the model and avoiding overfitting. This leads to approximately 4,500 samples for validation. Therefore, 85% of the training data is used for model training which consists of a large number of labeled samples. For this purpose, a different set of 5,000 samples is employed for purposes of testing the models. This testing set contains fake and real reviews. The dataset is balanced, and its division into training, validation, and test sets is reasonable enough to make a proper model assessment and serve as a reliable background for fake review detection.

4 Data Preprocessing

To clean the text data and prepare for ML and DL analysis several preprocessing operations were applied to transform the text into a structured format. The following are explained in detail below.

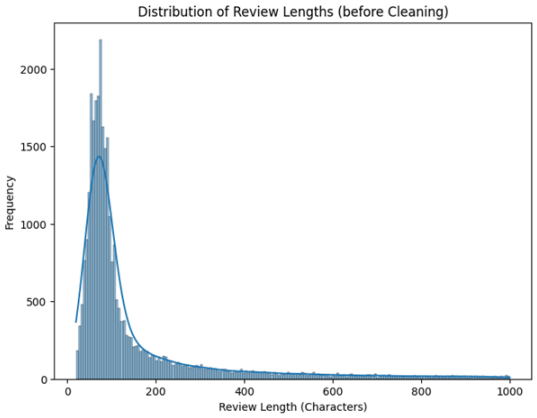


(a) Class distribution of labels (where OR and CG refers to original and computer generated reviews

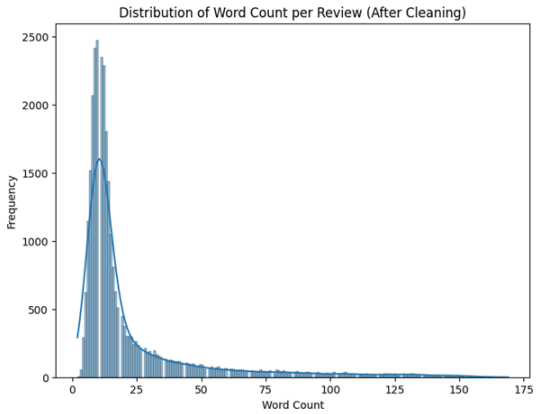


(b) Word cloud of dataset

Figure 1: Data visualisation Plots on the dataset



(a) Distribution text length before length filtering



(b) Distribution text length after length filtering

Figure 2: Distribution text length before and after length filtering

4.1 Checking for Null Values

As a initial step, the dataset was inspected for any possibility of null values in any of the features or any review. This step completes the dataset and avoids mistakes while training the model in the next steps of the process. In dataset there are no null values such that the dataset was ready for the next steps of preprocessing.

4.2 Text Cleaning

The next main step of text cleaning was to filter out any noise not required for the dataset or necessary to bring the input to a standardized format. Punctuation marks, symbols, and other characters were also excluded due to factors that can negatively impact the results of the analysis. Furthermore, all the text was transformed to lowercase so that all terms were handled as the same case regardless of the used case. Also, the preprocessed data excluded certain typical words which are referred to as stop words. Although these words occur quite often in the text, they do not convey great semantic information and their deletion contributes to the filtering of noise that distorts the dimensions of the data,

and makes the text more concise and easily computationally processed further.

4.3 Review Length Filtering

In order to include a simple measure of outliers in the data matrix, the cleaned review was also given as a feature, by calculating the length of each review. For purposes of information filtering, the number of characters in the reviews was employed as the criterion for determining the length of the review. Reviews with lengths less than 20 characters were considered and those with lengths greater than 1000 characters were also eliminated due to their tendential lack of content or structure. Using such thresholds, the dataset was then filtered to contain reviews of a particular length, to ensure that the results obtained are not skewed or biased by overly short or overly lengthy reviews. This step also helped in increasing the quality of data that was used in training as well as testing.

4.4 Text Vectorization

The text data was then cleaned filtered and converted to numerical form by using a Term Frequency-Inverse Document Frequency (TF-IDF) vectoriser. This approach portrays the relevance of words in every individual review to the overall frequency of the entire corpus. To improve the model process, the data was analyzed based on only a thousand most influential features. Furthermore, the vectorizer omitted English stopwords to get meaningful as well as domain-specific terms for developing the feature set. It also enabled the conversion of the textual data into numerical vectors, which is suitable for modeling to machine learning techniques nonetheless, it still maintains the semantic meaning of the text.

These preprocessing steps made the data clean, formatted, and ready for model training and testing hence developing a good model detecting fake reviews.

5 Baseline Models

The objective of this study is to compare the CNN-BiLSTM model for fake review detection against the following standard baseline models. These baselines simply help to compare the performance and the efficiency between the conventional algorithms of machine learning and the concept of deep learning. The chosen models are CNNs, LSTMs, RNNs, random forest, and naïve Bayes. These models are explained below, and why they have been included in this study is also explained.

5.1 Convolutional Neural Network (CNN)

CNNs (Convolutional Neural Networks) on the other hand are profoundly employed in different text classification because they are effective and efficient in ability to learn the localized features and patterns in textual data. In fake review detection, CNNs work well for the word level or n-gram level because it can often identify specific elements of what can be considered fake reviews, such as repeated phrases or keywords and odd word combinations. Although such local structures can be easily captured by CNNs, the relation between words in the large context of a review cannot be modeled by this network. This limitation is more critical in fake review detection as irregularities in

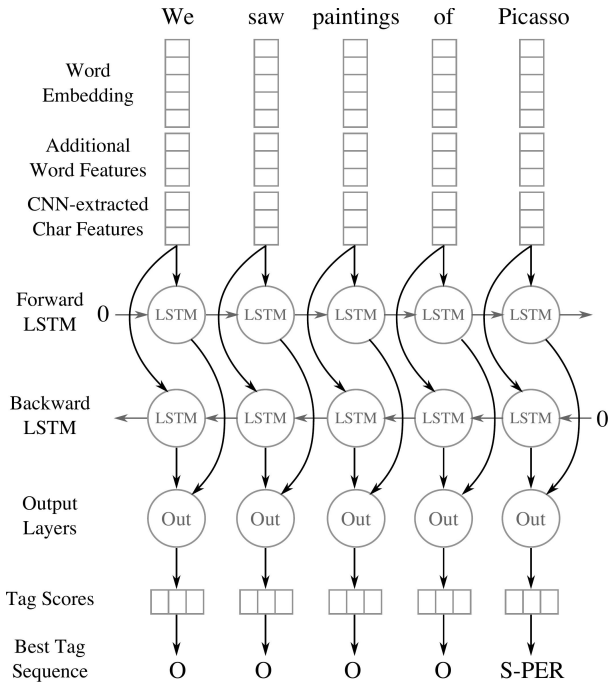


Figure 3: Basic Architecture of CNN+BiLSTM model

polarity shift, tone, and syntactic awkwardness usually extend across more extended sequences. Selecting CNN as the model with which other models such as CNN-BiLSTM can be compared and determine if the use of global context enhances the results of the local features.

5.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are recurrent neural networks (RNN) that are capable of processing the long-range relationships in the sequential data set. Fake reviews might contain subtle shifts in polarity or else include numerous and recurrent markings that can be hard to model with basic methods. Such patterns are especially detectable using LSTMs because they are capable of retaining information over long sequences. However, standard LSTMs, work in a unidirectional way, in the direction from

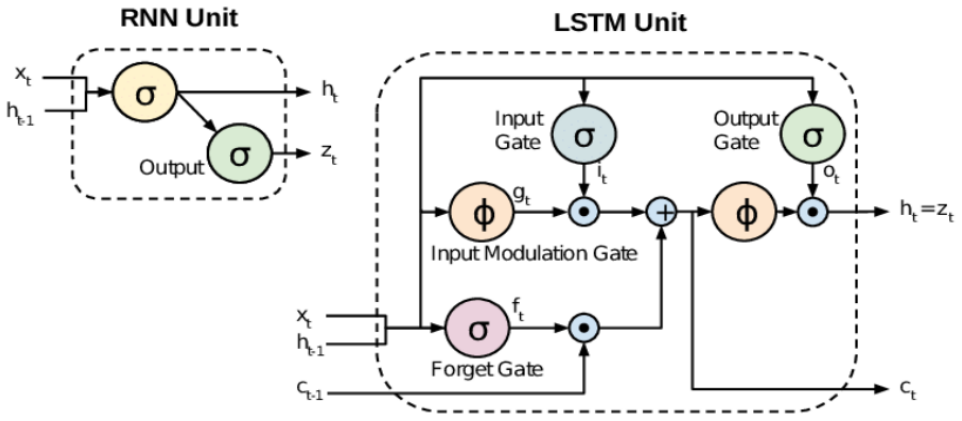


Figure 4: Basic architetcures of RNN and LSTM unit

past to present. This limitation limits them to capturing contextual relations fully in one direction of the review only. LSTM as a reference to evaluate the effects of bidirectional processing in comparison to bidirectional processing as used in the BiLSTM model and to analyze if unidirectional processing is already capable of identifying fake reviews.

5.3 Recurrent Neural Network (RNN)

There are other types of sequential models such as Recurrent Neural Networks (RNNs) in which the networks can handle text data since they retain some form of memory of the previous words in the sequence. Although, RNNs can mention some level of sequential dependency, which makes them appropriate to utilize in such tasks as sentiment analysis or fake review detection because in such a task, the sequence and flow of the words are significant. Nevertheless, the basic RNNs have some drawbacks, which are related to the vanishing gradient problem, and, therefore, are insufficient for capturing long-term dependencies. Moreover, the operation in RNN is unidirectional and only allows the flow of data through past contexts only. Selecting RNN as a starting point for comparison to show how it fares against other models such as BiLSTM, which works in two directions, providing a better perception of the general review.

5.4 Decision Tree Classifier

The Decision Tree classifier was chosen as it is simple, yet efficient for classification purposes. The hyperparameters that were varied in the Decision Tree model that chosen were `max_depth`, `min_samples_split`, and `criterion`. The parameter exploration for the `max_depth` was [None, 2, 10, 50, 100, 200, 500, 1000] and for `min_samples_split` values were [2, 5, 10, 20, 30, 50]. Further, the `criterion` parameter for selecting the function to measure the quality of a split was set with values as 'gini', and 'entropy'. Finally, the parameters were optimized to which the best values were `max_depth = None`, `min_samples_split = 5` and `criterion = 'entropy'`. This configuration enabled the tree to capture all the features of the data without distorting it and was efficient in fake review detection.

5.5 Random Forest Classifier

Random forest was chosen due to its flexibility in voting from multiple decision trees, which makes it suitable for high-dimension data. The hyperparameters that were adjusted for the Random Forest model included `n_estimators` and `max_depth`. The parameter ranges for `n_estimators` were set as [5, 10, 50, 100, 200, 300, 500] and for `max_depth` as [2, 3, 5, 10, 20, 30, 40, 50, 60]. Through experimentation, it was found that the best results were obtained when `n_estimators = 500` and `max_depth = 60`, which offered the highest accuracy. Such a configuration allowed the Random Forest model to identify feature interactions while minimizing overfitting.

5.6 Naive Bayes Classifier

The Naive Bayes classifier was adopted because of its suitability in text classification. The hyperparameter that was tuned for Naive Bayes was `var_smoothing` to prevent probabilities from overfitting the sparse features. The considered candidate values for

`var_smoothing` were as follows: $[1e-9, 1e-8, \ldots, 1]$ with logarithmic spacing. The optimum value of `var_smoothing` was 0.01 because it helped in smoothening the probabilities without compromising the accuracy. This was because Naive Bayes could work effectively in the fake review detection task although the algorithm cannot consider the sequential characteristics of the text.

5.7 Summary of Baseline Models

By incorporating these baseline models, the performance of the CNN-BiLSTM model is compared to that of fake review detection. All of these models have their strengths and weaknesses, and each is suitable for analyzing various strategies of this classification task. Compared with other types of machine learning models including decison tree, Random Forest, and Naive Bayes, these methods provide efficient and interpretable solutions, but the sequential data characteristics cannot be decoded. While CNN, LSTM, and RNN are more appropriate to follow the flow and contextual understanding of text they are not fully capable of capturing the bi-directional nature and long-range dependencies of fake reviews. The CNN-BiLSTM model, by combining the strengths of CNNs and LSTMs, is designed to overcome these challenges and provide a more accurate and comprehensive solution to fake review detection.

No.	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	Decision Tree	89.7	91.0	90.0	89.0
2	Naive Bayes	90.83	94.0	89.0	90.0
3	Random Forest	92.24	95.0	91.0	92.0
4	LSTM	94.75	93.0	95.0	94.0
5	RNN	94.38	92.0	96.0	94.0
6	CNN-BiLSTM	95.08	93.0	95.0	94.0

Table 1: Performance metrics of various models.

6 Results and Discussion

The performance of the various models in detecting fake reviews is summarized in Table 1 has all the metrics of accuracy, precision, recall, and F1 for each of the models.

6.1 Decision Tree

This Decision Tree model has a moderate accuracy of 89.7% and these results show the simplicity of the Decision Tree model. Decision Trees are known to overfit and are unable to capture the interactions in data. It only achieves a high accuracy of 91.0% in the cases of real reviews but has a low recall of 90.0% for fake reviews. Hence the lower F1 score of 89.0% evidence that the model fails to balance both precision and recall.

6.2 Naive Bayes

Naive Bayes has slightly less error than Decision Tree, and it gives an accuracy of 90.83%. This is especially impressive for precision (94.0%), which means the developed model is

good at identifying non-fake reviews, but recall (89.0%) is lower, suggesting some fake reviews will not be identified. This performance disparity is again visible in the F1 score of 90.0%. Naive Bayes classifier works on simple probability assumptions and goes by the principle that all the features are equally unrelated. This independence assumption may not always be true in the case of text datasets because words and phrases are dependent. For fake review detection, where subtle language features are important, Naive Bayes may fail to capture complex relations between the words hence the lower recall and a comparatively poor F1 score.

6.3 Random Forest

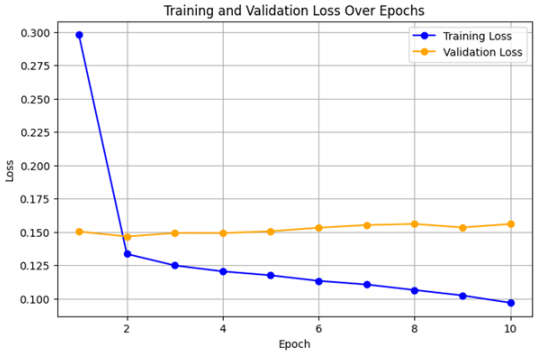
Random Forest has a reasonable accuracy of 92.24% and the precision of a model is 95.0 %. The model also has a good recall rate of 91.0% which means that the model can perform well in balancing both false positives and false negatives. Random Forest is a technique where you have several decision trees, this reduces overfitting, which is a problem that may affect a single decision tree. This includes a better generalization which is very crucial when it comes to identifying fake reviews from various sources. However, while Random Forest has a good showing, it cannot extract temporal and contextual correlation in text which is critically important for interpreting the complexities of fake reviews. This limitation is well illustrated in its performance, the model's recall (91.0%) is lower than that of more complex models such as LSTM, RNN, and CNN-BiLSTM which are better placed to capture such dependencies.

6.4 LSTM (Long Short-Term Memory)

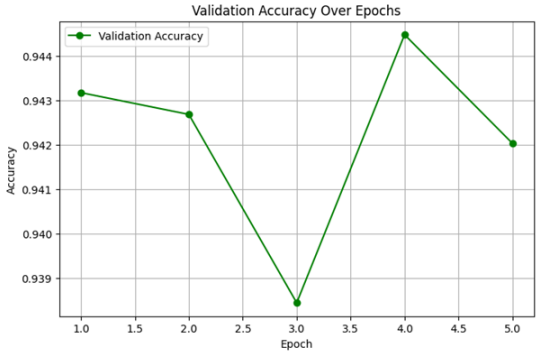
In the LSTM model, the accuracy is 94.75%, the precision value is 93.0%, and the recall value is 95.0. LSTM can keep the long dependency of sequences, so it is effective in catching the contextual meaning of words in the review and helps to detect fake reviews which may be not significantly different from the original one but have fake meanings. Although LSTM is very effective, it is still not capable of utilizing the local features of text such as n-grams or some specific word phrases which are also helpful in detecting fake reviews. This is where combining LSTM with CNN provides an added advantage, as described in the later sections of this paper.

6.5 RNN

The RNN model is as effective as LSTM with an accuracy of 94.38%, although the RNN has 96.0% recall and 92.0% precision. RNNs, as well as LSTMs, are usually used to handle sequential data. However, standard implementations of recurrent neural networks have certain drawbacks such as vanishing gradients; therefore they are inferior to LSTMs for learning long-term dependencies. However, it can be noted that even in this experiment, the RNN model driving this experiment can capture the sequential nature of the text better and thus it gets a higher recall than Decision Trees, Naive Bayes, and Random Forest. However, the lower precision implies that RNNs might learn some sort of pattern in the data set and hence the false positives.



(a) Training and validation loss plot for RNN



(b) Validation accuracy over the epochs for RNN

Figure 5: Training plots of RNN

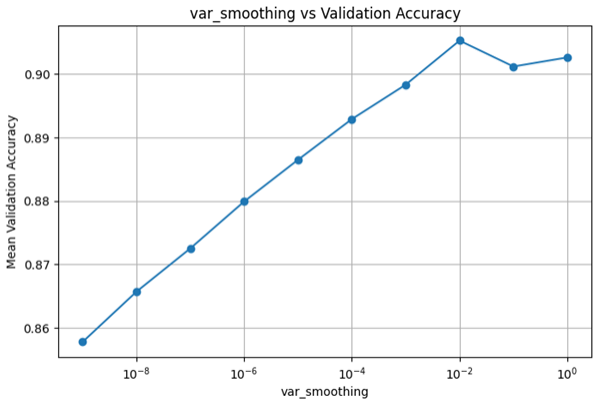


Figure 6: Validation Accuracy vs var_smoothing in Naive Bayes model

6.6 CNN-BiLSTM

CNN-BiLSTM has performed efficiently compared to other models with the highest accuracy of **95.08%**. It achieves a precision of **93.0%**, recall of **95.0%**, and F1 score of **94.0%**. The combined architecture of CNN and BiLSTM is particularly suitable for fake review detection, as it can capture both local and global textual patterns.

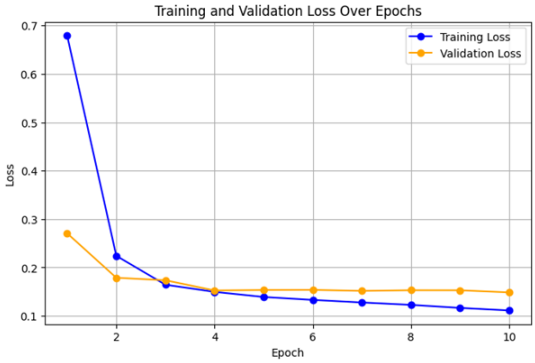
6.6.1 Convolutional Neural Network (CNN)

CNNs are the best when it comes to finding local patterns, for instance, n-grams, recurring phrases, and particular sequences of words. These are the features significant for fake detection since such reviews tend to have slight differences in the repeated cues. This layer scans the text in small windows and extracts features which in turn will help the BiLSTM layer. This makes it possible for CNN to identify features from the text that are otherwise hidden in the raw input but which are very suggestive of fake reviews.

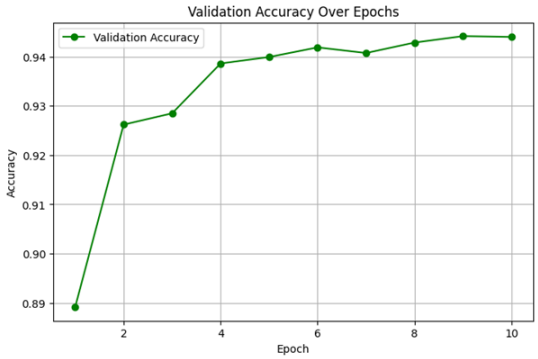
6.6.2 Bi-directional Long Short-Term Memory (BiLSTM)

The inclusion of BiLSTM improves the model by processing the input text forward and backward to make the determined tags. This means that the model can pick up context information from both the previous and the next elements in the sequence, which will

help the model to have a better understanding of the context of the review. Previous textual context is crucial in the fake review detection together with the following textual context, for instance, a shift from positive to negative. One such benefit of bidirectional LSTM is that it allows the model to understand the structure that fake reviews take, even if it is at some point later in the review.

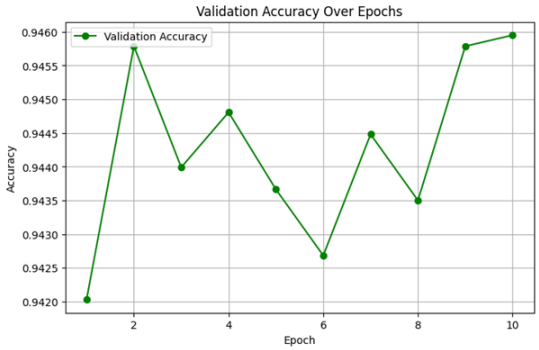


(a) Training and validation loss plot for CNN+BiLSTM



(b) Validation accuracy over the epochs for CNN+BiLSTM

(c) Training and validation loss plot for LSTM



(d) Validation accuracy over the epochs for LSTM

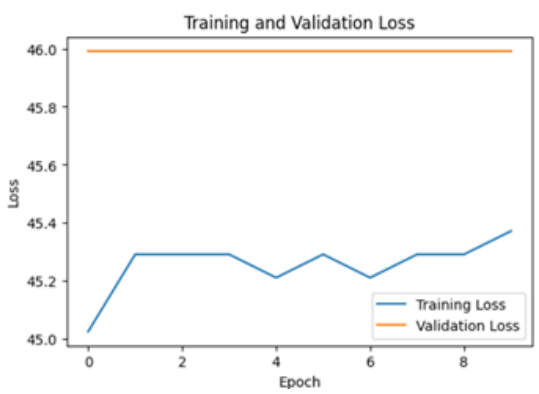
Figure 7: Training plots of CNN+BiLSTM and LSTM models

6.7 Why CNN-BiLSTM Performs Better

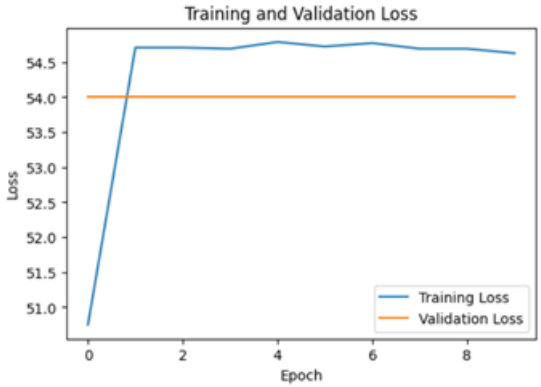
CNN-BiLSTM model has combined advantages of convolutional and sequential learning, which is more suitable for fake review detection. CNNs are effective for distinguishing local information patterns, including certain word associations, indicating that a particular review is fake, whereas BiLSTMs are more effective at capturing sequential information characteristics of the text. This makes the CNN-BiLSTM model ideal for fake review detection where not only local features such as fragment combination, in terms of words, are relevant but also global features such as the sentiment of a fragment, and its structure.

In addition, as the CNN can capture the local dependencies and the BiLSTM can capture the long-distance contexts, the proposed CNN-BiLSTM model can understand more linguistic features that distinguish a fake review from a real one. Some of the models that we compared with CNN-BiLSTM include Decision Trees, Naive Bayes, and Random

Forests, and these are incapable of modeling the complex patterns of the text thus their low performance compared to CNN-BiLSTM. As for LSTM and RNN models, similar to the previous experiment, they are also effective in modeling sequential dependencies In this case, however, CNN-BiLSTM adds an extra level of difficulty in modeling both local and global dependencies in text data, and that makes the difference.



(a) Training and validation loss plot for CNN+BiLSTM with learning rate 0.01

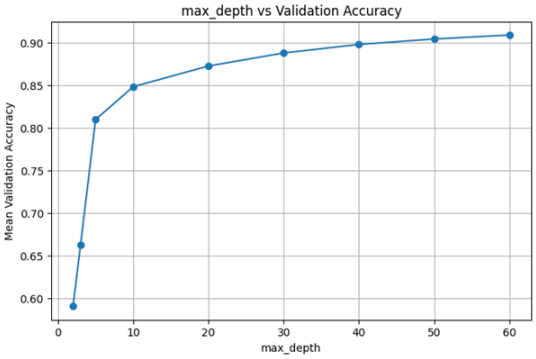


(b) Training and validation loss plot for CNN+BiLSTM with learning rate 0.1

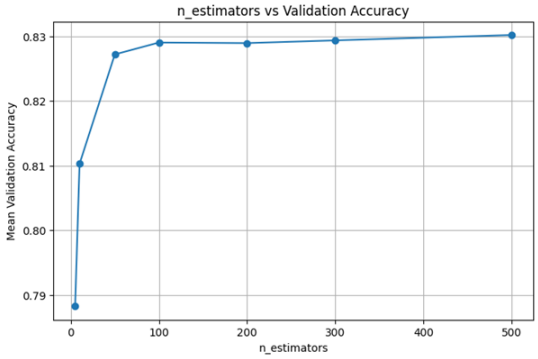
Figure 8: For CNN+BiLSTM, various learning were tried, where with learning rates 0.01 and 0.1 models are unable to converge, and with learning rate of 0.001 suited well and able to converge. Results reported in Table 1 for the deep learning models are with 0.001 learning rate and loss plot can be seen in figure 7a

Therefore, the proposed **CNN-BiLSTM model** is the best among all other baseline models as it can handle both local and global features in the text. CNN's feature extraction power and BiLSTM's ability to handle sequential dependencies make it particularly effective for detecting fake reviews, which often contain intricate patterns that simpler models like Decision Trees or Naive Bayes cannot fully exploit. As a result, CNN-BiLSTM is the best model for this task, providing a robust and accurate solution for fake review detection.

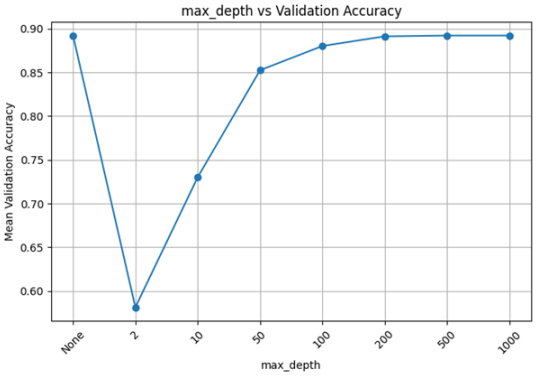
For the deep learning models, optimization was important for the learning rates because they have a strong impact on the models while training. In the experiments with learning rates of 0.1, 0.01, and 0.001 the best result was achieved with the rate of 0.001. The set of validation was 10% of the training data to check the performance and select the ideal model for avoiding overtraining. The selected models were trained and tested on the test dataset proving their efficacy to generalize well. With this systematic approach to tuning, it becomes clear the importance of appropriately selecting hyperparameters such as learning rates in the delivery of optimal performance for deep learning structures.



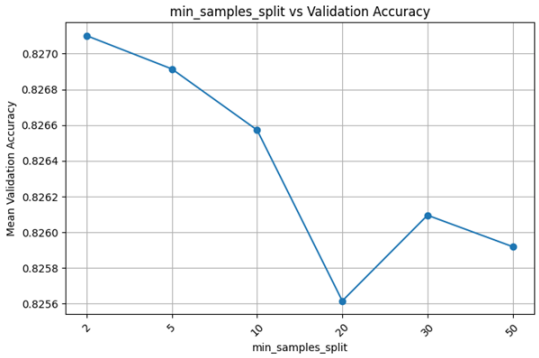
(a) In Random Forest Validation accuracy vs Max depth



(b) In Random Forest Validation accuracy vs n_estimators



(c) In decison tree, Validation accuracy vs Max depth



(d) In decison tree Validation accuracy vs Min samples per split

Figure 9: Hyper-Parameter tuning for Random Forest and Decision Tree

7 Conclusion and Future Work

7.1 Conclusion

This research successfully achieved the goal of using machine learning and deep learning for the identification of fake reviews in ecommerce platforms. Using a large, original, and synthetic set of reviews that are genearted using OLAAMA, the study systematically evaluated different models including the Decision Trees, Random Forests, Gaussian Naive Bayes, and more complex deep learning models CNN-BiLSTM. As it was expected, the CNN BiLSTM model demonstrated the highest accuracy as it provided both the local features and the contextual information. This study expands the knowledge on the use of a combination of deep learning approaches to solve text classification issues in online shopping to increase customer confidence and satisfaction.

7.2 Future Work

The limitations highlighted in this study should be the focus of future work, including the requirement of large labeled data and domain transferability. Further research into the use of unsupervised and semi-supervised learning could go a long way toward reducing the importance of labeling, providing models the ability to change as the nature of reviews changes. Furthermore, the introduction of transfer learning approaches with other

established methods such as BERT or GPT may enhance the chance to distinguish rather sophisticated and context-based fake comments. Another future work idea is to enlarge the dataset to incorporate multiple-language reviews, as well as employ reinforcement learning for retraining on new forms of deception to improve the efficiency of fake review detection systems.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023). Gpt-4 technical report, *arXiv preprint arXiv:2303.08774* .

Alsharif, N. (2022). Fake opinion detection in an e-commerce business based on a long-short memory algorithm, *Soft Computing* **26**(16): 7847–7854.

Alsubari, S. N., Deshmukh, S. N., Aldhyani, T. H., Al Nefaie, A. H. and Alrasheedi, M. (2022). Data analytics for the identification of fake reviews using supervised learning, *Computers, Materials & Continua* **70**(2): 3189–3204.

Alsubari, S. N., Deshmukh, S. N., Aldhyani, T. H., Al Nefaie, A. H. and Alrasheedi, M. (2023). Rule-based classifiers for identifying fake reviews in e-commerce: A deep learning system, *Fuzzy, Rough and Intuitionistic Fuzzy Set Approaches for Data Handling: Theory and Applications*, Springer, pp. 257–276.

Baishya, D., Deka, J. J., Dey, G. and Singh, P. K. (2021). Safer: Sentiment analysis-based fake review detection in e-commerce using deep learning, *SN Computer Science* **2**(6): 479.

Bathla, G. and Kumar, A. (2021). Opinion spam detection using deep learning, *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, pp. 1160–1164.

Fusilier, D. W., Montes-y Gómez, M., Rosso, P. and Cabrera, R. G. (2015). Detecting positive and negative deceptive opinions using pu-learning, *Information Processing & Management* **51**(4): 433–443.

Li, J., Ott, M., Cardie, C. and Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam, *Proceedings of ACL*, pp. 1566–1576.

Liu, M., Shang, Y., Yue, Q. and Zhou, J. (2021). Detecting fake reviews using multidimensional representations with fine-grained aspects plan, *IEEE Access* **9**: 3765–3773.

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S. et al. (2020). Language models are few-shot learners, *arXiv preprint arXiv:2005.14165* **1**.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A. (2023). A comprehensive overview of large language models, *arXiv preprint arXiv:2307.06435* .

Pavlinek, M. and Podgorelec, V. (2017). Text classification method based on self-training and lda topic models, *Expert Systems with Applications* **80**: 83–93.

Rayana, S. and Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata, *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Sumathi, V. P., Pudhiyavan, S. M., Saran, M. and Kumar, V. N. (2021). Fake review detection of e-commerce electronic products using machine learning techniques, *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, IEEE, pp. 1–5.

Thuy, D. T. T., Thuy, L. T. M., Bach, N. C., Duc, T. T., Bach, H. G. and Cuong, D. D. (2024). Designing a deep learning-based application for detecting fake online reviews, *Engineering Applications of Artificial Intelligence* **134**: 108708.

Tufail, H., Ashraf, M. U., Alsubhi, K. and Aljahdali, H. M. (2022). The effect of fake reviews on e-commerce during and after covid-19 pandemic: Skl-based fake reviews detection, *IEEE Access* **10**: 25555–25564.

Yelundur, A. R., Chaoji, V. and Mishra, B. (2019). Detection of review abuse via semi-supervised binary multi-target tensor decomposition, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2134–2144.

Zhang, H., Zhang, C. and Liu, X. (2023). Fake review detection via deep learning techniques: A comprehensive survey, *Journal of Computational Science* **69**: 101348.

Zhang, X., Sun, Y., Zhang, H., Zhang, Z. and Li, Z. (2020). Detecting fake reviews using an ensemble learning approach, *Applied Sciences* **10**(20): 7149.