National College of Ireland

Identify Counterfeit Product Reviews or AI Text by Bots on E-commerce Websites

MSc Research Project
Data Analytics

Avis Massey
Student ID: x21199752

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Avis Massey
Student ID:	x21199752
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Aaloka Anant
Submission Due Date:	14/12/2023
Project Title:	Identify Counterfeit Product Reviews or AI Text by Bots on E-commerce Websites
Word Count:	8917
Page Count:	24

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:	
Date:	31st January 2024

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):	

Identify Counterfeit Product Reviews or AI Text by Bots on E-commerce Websites

Avis Massey
x21199752

Abstract

Digital technologies are becoming a part of our day-to-day decision-making. Potential buyers explore reviews and experiences from the current users of the products and services on various online platforms that influence their buying decisions. Being said that, the use of these tools has been widely used to manipulate the perceptions of consumers and influence the buying decision in favour of the entities, making fake reviews a critical challenge. Potential buyers are influenced by the popular service providing websites and e-commerce platforms through encouraged user feedback. With the rising dependency on product feedback, it becomes paramount for the customers to be able to identify the genuine feedback from the pool of false reviews. To vanquish this challenge for potential buyers, we move to advanced machine learning and deep learning techniques, applying models like LSTM, CNN, and SVM for deep learning and BERT, Roberta, Albert, and DistilBERT for transformer models. The crucial component of consumer decision-making has shifted to online reviews, which people share based on their actual experiences. In all appearances, the increase in exploitation of technology is leading to misguided consumer choices by generating spam and fake reviews to either boost or undermine a business. Marketers can utilise this analysis to ensure a transparent and trustworthy online marketplace by using tailored strategies and to customer preferences. We evaluate models' performance based on accuracy and weighted F1-score, which demonstrates the superior capabilities of a model in detecting false reviews.

1 Introduction

In the contemporary consumer landscape, online reviews have transformed into influential decision-making tools, significantly impacting perceptions and shaping market dynamics.Positive evaluations build emotional trust between consumers and brands, influencing purchasing decisions beyond factual knowledge. They convey trust and confidence, increasing product or service uptake. Negative ratings alert consumers to potential hazards and encourage risk-aversion. Positive or negative reviews' emotional tone strongly influences buyer impressions. Positive emotions boost product desirability and buying intent, while negative emotions lower it. Businesses are realising they need to actively manage their digital reputation as online reviews become more important. This includes using favourable reviews and addressing bad feedback. In the competitive market, actively engaging with customers to lessen bad reviews shows a dedication to customer happiness and is crucial to a brand's success. Businesses must navigate and effectively respond to internet reviews' emotional and informational dynamics to succeed in the new customer

marketplace. Online reviews are crucial to consumer perception and decision-making. Positive and negative evaluations' emotional nuances affect consumer trust and confidence. Positive ratings boost consumer happiness and brand image. Negative reviews may dissuade purchasers owing to perceived risks or concerns. Online reviews from verified buyers or trusted sources are more credible. Reviews' influence depends on their relevance, especially in meeting potential purchasers' needs. To maintain a good online reputation, businesses must actively manage these aspects. Reviews' trustworthiness and reliability affect brand perception. Hedonic products are reviewed for emotional gratification, while utilitarian ones are reviewed for practicality. Businesses can modify their approach to different product categories and consumer expectations by recognising these distinctions. Businesses recognise the importance of online reviews because they actively influence consumer behaviour. Positive reviews boost brand reputation and consumer trust. Effectively handling unfavourable evaluations might reveal areas for development. Business use text mining to gain insights from the enormous pool of online reviews. Review analysis helps uncover consumer opinions, preferences, and pain areas. Text mining helps organisations adjust products and services to changing client needs and use online input for business growth. Businesses encourage and monitor online reviews and use advanced analytics to gain meaningful insights from consumer input. In the last few decades, the rise in spam reviews on websites has caused a lot of worry, which is why experts have built strong systems that can accurately spot fake reviews. The number of fake reviews went from 5% in 2006 to 20% in 2013. This made it necessary to look into different methods, such as machine learning and deep learning models. Jindal and Liu's early study in 2007 found problems that led to more research and better methods. Ott and his colleagues created a benchmark dataset called the "gold standard" in 2018. It was able to read false reviews 86% of the time, which is very good. To learn more about fake user reviews, academics have looked into machine learning models, deep learning-based models, and transformer-based models, with a focus on deep learning studies. A lot of focus has been paid to Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for finding false reviews in machine learning and deep learning. Specific instance accuracy of 94% have been reached with LSTM, which is known for recording temporal dependencies. It has been emphasised that to improve accuracy, contextual information like user behaviour, social network structure, and time dynamics should be added to deep learning models. Convolutional Neural Networks (CNNs) are known for keeping data safe and keeping people from getting false information. Using different approaches, like Particle Swarm Optimisation (PSO) and hybrid CNN-LSTM models, has made false review identification much more accurate. Support Vector Machine (SVM) has been popular, with an F1-score of 80% and a recall of 81%. In the beginning, Logistic Regression was the most popular method. Even if SVM works, it might have trouble with reviews that are neutral or positive. This shows that we need more advanced natural language processing methods. Models try to be as accurate as possible so that internet reviews can be trusted and so that buyers and sellers don't have to worry about losing money. Adding semantic analysis to SVM and NB algorithms has fixed problems and made them more accurate. This concludes that using models like SVM, LSTM, and CNN is a strong way to make online review systems more reliable. Some of the best text classification models right now are transformer-based pre-trained models like BERT, RoBERTa, ALBERT, and DistilBERT. Their work to find fake reviews is still in its early stages, but it looks like it will be interesting. Some well-known deep learning models, like CNN-LSTM by Alsubari et al. (2021) (Frontiers;

2021) and the FABC hybrid model by Jacob and Selvi Rajendran (2022) (Frontiers; 2021), use both CNN and LSTM to get better contextual knowledge. But these deep learning models have trouble with being able to handle big datasets and not being able to do computations in parallel. To solve this problem, transformer models like BERT, RoBERTa, and DISTILBERT have been created. RoBERTa has shown particularly good results (Gupta et al.; 2021). To sum up, standard machine learning is giving way to more advanced deep learning and transformer-based methods for finding fake reviews. These new methods aim to be more accurate, scale-able, and useful in all situations.

1.1 Research Question

How to detect and identify counterfeit product reviews authored by automated bots on e-commerce platforms by utilizing various machine learning techniques and methodologies?

1.2 Research Objective

This research aims to make a significant contribution by developing a resource efficient and effective model for identification if false and counterfeit reviews across multiple e-commerce platforms by employing multiple transfer learning techniques and models. The objective and contribution are as follows:

1. Thorough investigation and analysis of the existing research on false review identification.

2. Data cleaning and pre-processing to align with input parameter requirements for deep and transfer learning models

3. Implementing LSTM, CNN , SVM and BERT, RoBERTa, ALBERT, and DistilBERT models for the false review classification task across various dataset samples, while achieving meaningful results with a small labeled dataset and limited computational resources.

4. Evaluating and comparing the performance of all implemented models using graphical interpretation, considering the accuracy and f1-score as the computational metrics.

5. Testing the pre-trained models on a new dataset to evaluate the performance

2 Related Work

In the last decades, there has been substantial rise in spam reviews on online platforms. Due to this surge, researches are inclined towards developing a robust system which will be capable of accurately identify false reviews. According to a report, in 2006 the rate of false reviews was 5% which saw an increase to 20% in the year 2013 (Hai et al.; 2016). Researches have explored different methods which include machine learning models and deep learning models. These points make it cardinal to analyze it with different advancements in the field of machine learning (Kim et al.; 2021). This section gives an overview of the literature on various approaches which helps us to explore detection of false reviews.

The initial research in this domain was done by S.N. Jindal and B. Liu in 2007, which can be referred to in 'Review Spam Detection' studies (Jindal and Liu; 2007). In their subsequent work (Jindal and Liu; 2008), they had expressed multiple challenges in detection of false reviews, there was an emphasis on additional research and enhancement in the methods of detection. Furthurmore, M. Ott and collaborators introduced a 'gold standard' dataset in 2018, which used to serve as a benchmark and were able to interpret false reviews with an accuracy of 86% (Ott; 2018) (Ott et al.; 2013). The online counterfeit reviews is characterized by diverse and consize textual content. To apply existing techniques of machine learning to such complex data will be challenging, which may lead to sub optimal accuracies in the process of identification. In the recent years we have seen a surge in the realm of Deep learning studies dedicated to enhance the comprehension of false user reviews. (Taşağal and Uçar; 2018). In this section, we explore the literature reviews encompassing various techniques, including machine learning models, deep learning-based models and transformer based models.

2.1 Machine Learning and Deep Learning Models

In the recent years there has been a noticeable surge in interest related to deep learning (DL) methods. Mainly in those which employ Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for the detection of false reviews. LSTM being a RNN, it is adpt in capturing the temporal dependencies in the sequential data, we can process the text-data which is intricate and have prolonged dependencies as found in false reviews making it suitable for rendering. In a paper authored by (Varlamis et al.; 2021), they introduced an LSTM-based model that was able to achieve an accuracy of 94%. In another paper which is titled "A Comprehensive Review on Fake News Detection with Deep Learning," the authors in the research did an extensive examination of the features used in DL models for the detection of false reviews. They emphasized the importance of encompassing visual features, textual features and social context features. (K. et al.; 2019) informs us about the significance of contextual information such as user behaviour, social network structure and temporal dynamics into the DL models to give a boost to the accuracy. (Wang et al.; 2018) in their developed a model for detecting false reviews by employing LSTM, RNN. The model uses textual features for training and the evaluation is conducted on the dataset which is scraped from web pages which are accessible in Taiwan. The architecture of the model comprises on an input layer, LSTM layer, and output layer with the hidden layers included for dimension reduction. It was reported that the LSTM models performed better than the baseline SVM model. There are primarily two challenges, memorization-extreme learning and the issue of vanishing gradient. To address the problem of memorization arising from the proliferation of parameters in the hidden layers we apply the dropout method which was proposed by (Srivastava et al.; 2014). For Natural Language Processing (NLP), RNN is a commonly used method which is capable of predicting the next word based on preceding words in a given text. The RNN utilizes a backpropogation algorithm, but again while during this process the gradient may lead to zero arising the problem. To resolve the problem LSTM architecture was introduced (Hochreiter and Schmidhuber; 1997) which comprises of the input, output, sigmoid (or tangent), and forget-doors, which are distinctive components not present in the RNN algorithm. LSTM models have proven to be effective in detection of false reviews. The advantages include its ability to capture long-term dependencies within the data, this will help in indentification of intricate pat-

terns and relationships. In the study given by (Vyas et al.; 2021) for detection of false reviews the CNN – LSTM model proposed to be of superior performance. Authentic labelled dataset for LSTM based method specifically tailored for the detection of false news proves to be outperforming different models. (Desai et al.; 2023)In conclusion, leveraging LSTM models for fake review detection offers a robust approach to enhance the credibility of online review system.

The Convolutional Neural Network (CNN) has garnered imperative attention due to its credibility in holding data integrity and safeguarding users from deceptive information. There have been multiple studies which has employed diverse models and methods which is aimed at enhancing the precision in detection of false reviews. The Particle Swarm Optimization (PSO) algorithm which is known to identify critical features of a text data which results in high accuracy during the training and testing phases (Deshai and Rao; 2023). The BSTC model which combines a pre trained model with CNN is able to learn document level data which is helpful in false review detection. We also see a hybrid approach of CNN-LSTM model leveraging the networks of both models also significantly enhances the accuracy in false review detection (Deshai and Rao; 2023). In a study where the LSTM, BILSTM, and CNN-LSTM models were trained to detect reviews wherein the LSTM model was able to achieve an impressive accuracy of 87% (Taşağal and Uçar; 2018). These models underscore the potential of CNN in uncovering the false review detection by extracting meaningful features from textual data, although it is crucial to consider variations in performance based on specific datasets and potential over-fitting in training data (Deshai and Rao; 2023)

Several research papers and articles have explored into the application of Support Vector Machine (SVM) for identification of false reviews, Notable using the SVM algorithm reported in the research by (Tellawar et al.; 2023) were able to achieve recall and F1-score of 80%, in an another research the SVM classifier was able to achieve an accuracy rate of 87.81% in the identification of false reviews (Mir et al.; 2023). But there are few limitations to the model where in it tends to make more errors when dealing with false reviews expressed in a neutral or positive tone and those utilizing common language. The overall motive of the research is to achieve an optimal accuracy, a critical aspect for upholding the e credibility of online reviews and mitigating financial risks for both consumers and sellers (Tellawar et al.; 2023) (Mir et al.; 2023). In the early research of false review detection by (Jindal and Liu; 2008), the study conducted had employed models such as Bayesian algorithm, SVM, and Logistic regression where in Logistic Regression dominated the performance among these algorithms, but the limitation in the study faced was distinguishing between false and genuine reviews which led to considerations of duplicates as false. In a similar study (Hassan and Islam; 2020) they had utilized multiple ML algorithms which included logistic regression, SVM, and the Naïve Bayes classifier. In the study the SVM classifier was able to achieve an accuracy of 88.75% but again there was a drawback due to relatively small dataset used for analysis. In a study proposed by (Khan et al.; 2021) Models like SVM, Naïve Bayes, Decision tree, and Logistic regression have been used among them, the SVM performed better than the other models. But the limitation with the model was that the dataset was imbalanced and TF IDF was the only feature extraction technique used. The given method was to identify the false reviews in the domain of social media, the integrated algorithm was suppose to increase the accuracy and applicability in the same domain. The combination of SVM and NV was effective in minimization of the counterfeit reviews and maximize the balance detection rate. Furthermore adding semantics analysis to the SVM and NB

classifier addressed a limitation of assuming independent feature whereby improving the accuracy of SVM by focusing on informative subspace of the feature spaces. Hence using SVM classifier model can be a crucial model to achieve a desired accuracy.

2.2 Transformer and Large Language Models

The introduction of transformer models, particularly BERT (Bidirectional Encoder Representation) has revolutionized natural language processing (NLP) tasks, offering a robust foundation for various applications. Different capabilities of BERT and other transformer models have been explored across diverse domains. (Devlin et al.; 2019) introduced BERT as a pre-trained transformer model capable of bidirectional analysis of text content. The model employs a Masked Language Model (MLM) to effectively achieve bi-directionality, which allows it to consider context from both preceding and succeeding words. BERT's pre-trained nature enables its application in various downstream NLP tasks without extensive task-specific training. Various researchers have extensively employed BERT model for sentiment analysis and classification task. In a study done by (Abdul et al.; 2019), the model was used to determine the polarity of reviews in their IMDB dataset, they were able to achieve an F1-score of 89%. The model had ability to understand contextual errors in language which contributed in distinguishing sentiment. But the challenge in their model was feature extraction from shorter texts. To address this issue, (Hu et al.; 2022)proposed a methodology using BERT model which analyzed mental features of reviewers which used to enhance the extraction of meaningful features from short texts. It was proved that the incorporation of mental features was able to improve the accuracy, particularly in predicting fake reviews within shorter texts. While BERT demonstrates remarkable performance, its inference accuracy can be influenced by different domains. It is essential to consider the context and nature of the data when applying BERT across diverse domains (Hu et al.; 2022). This factor emphasizes the importance of understanding the model's adaptability and potential limitations. Though BERT demonstrates excellent performance, its accuracy can influence different domains. The research suggest that the context and nature of the data when you apply BERT across diverse domains (Hu et al.; 2022). In a further study, (Shan et al.; 2021) had employed transformer-based models to identify misinformation related to the COVID-19 pandemic across social media platforms and the findings indicated that the employing domain-specific language models led to enhanced performance in the context of sequence classification tasks. These factors proves the importance of understanding the models adaptability and potential limitation. When comparing BERT with other traditional models, such as SVM, we can highlight the superiority of BERT in capturing complex linguistic structures. The bidirectional ability if BERT enables it to outperform other modes in the tasks requiring a nuanced understanding of language (Hu et al.; 2022).The pre-trained nature of BERT, coupled with its bidirectional capabilities, makes it particularly effective in capturing intricate linguistic patterns and addressing challenges such as feature extraction from short texts which makes it an ideal model to employ for our project. In the recent years of advancements with the models particulary RoBERTa, BERT, ALBERT, and DistilBERT, all rooted in transformer architectures. (Liu et al.; 2019) introduced the model RoBERTa as an enhanced version of BERT, which emphasizes the increased pre-trained data for improved accuracy. In a study the RoBERTa model exhibited superior performance over BERT, which was able to achieving optimal results in datasets like SQUAD and GLUE with minimal fine-tuning. Considering Its application in false review detection, employing

this transfer learning, has yielded promising outcomes, which resulted in outperforming baseline models in accuracy and weighted F1-score, as demonstrated in various studies. But there has been few limitation with the RoBERTa model such as the time-consuming fine-tuning process and the absence of a 10-fold cross-validation approach in some experiments (Gupta et al.; 2021). Despite these limitations, the research has consistently showcased the advantages of RoBERTa over traditional machine learning models wherein it has notably surpassed the SVM model in accuracy and F1-score metrics which makes it an model for us to employ.

DistilBERT, another transformer-based pre-trained model, has also entered into the domain of fake review detection through transfer learning techniques. DistilBERT's competitive performance on the NLP tasks is exceptional with the output achieving an accuracy of 68% with a weighted F1-score of 0.68, which is highlighted in studies comparing its effectiveness with other transformer models like BERT, RoBERTa, and ALBERT. This studies makes it evident to employthe potential of transfer learning in enhancing false review detection across various pre-trained models. In the pursuit of refining transformer models to employ it for the process of revirew detectin, (He et al.; 2020) introduced DeBERTa,which had surpassed the BERT and RoBERTa models by incorporating disentangled attention techniques and has also introduced an improved masked decoder. While not directly applied to false review detection which gives us a motive to explore it in the domain, DeBERTa models advancements contribute to the ongoing evolution of transformer models for diverse natural language processing tasks. We also explore the ALBERT model which is another transformer-based pre-trained model utilized in the false review detection task. Comparative studies involving BERT, RoBERTa, ALBERT, and DistilBERT models have consistently demonstrated the superior performance of RoBERTa in detecting fake reviews but our analysis says otherwise. Deep learning approaches which incorporate the ALBERT model , have achieved state-of-the-art results in false review detection, which we will be exploring to enhance and contribute in the advancement of the modle. In conclusion by the integration of transformer-based pre-trained modelslike RoBERTa, DistilBERT, and ALBERT, has seen a significant advancement in the domain of false review detection. These models, with their unique architectures and transfer learning capabilities has showcased promising outcomes, but in considerations such as fine-tuning complexities and varying dataset sizes.

3 Methodology

The research aims to develop effective models of deep learning and transformer learning for the detection of false online reviews. To conduct a data science project we employ a well structured methodology named Cross-Industry Standard Process for Data Mining (CRISP-DM).

3.1 Business Understanding

In the contemporary landscape of consumerism, online reviews have evolved beyond mere informative tools, becoming a cornerstone in the decision-making process. There is a considerable influence on consumer behavior, shaping perceptions and perceptions of the marketplace from the dynamic source of real-world experiences. Consumers tend to establish emotional trust on products and brands based on positive reviews, this trust between the brand and consumer extends beyond the factual information due to the

positive reviews. Often potential buyers purchasing decisions are swayed in a positive direction when encountered by favorable experiences shared by others creating a sense of reliability and confidence in the product or services. Multiple e-commerce website have these types of reviews flooded, As these reviews can be essential in descision making it becomes essential to check for the effectiveness and credibility. Below are enumerated different types of false reviews practices.

1. **Consumer Protection:** By distinguishing the false reviews with the false ones, it directly contributes to the consumer protection, it makes sure that the consumers are not misled by counterfeit reviews which may lead to poor purchasing decisions. Secondly, protecting consumers from these deceptive practices will increase their trust in e-commerce platforms and also promotes healthier online marketplace.

2. **Brand Protection:** Brands often suffer from false negative reviews which directly impacts their reputation. This research aids in brand protection by identifying and filtering out these false reviews. This practice is imperative so the integrity and reputation of the brand is only judge by genuine customer feedback.

3. **Combatting Sock Puppeting:** "Sock Puppeting" is a term wherein individuals or entities create a fake ID to write positive reviews for themselves or negative reviews for the competitor brand at times using bot or AI, which is a significant issue in retail market. Hence this research helps in identification of such fraudulent activities and maintains a fair competition.

Hence it is imperative to employ machine learning as a tool to descern counterfeit reviews with the true ones.

3.2 Data Understanding

The dataset included in the study is equally divided between authentic Amazon customer reviews and synthetically generated fake reviews which are generated by using GPT and ULFIT. Numerically the data is divided in 2.5K genuine reviews and 2.5K counterfeit reviews. The data was sourced from (Salminen et al.; 2022) which is a general available dataset. The dataset becomes ideal for doing comprehensive analysis on the models.

3.3 Data Processing and Modelling

The first steps in data-processing involves before using it for the models. These process involves the removal of null and duplicate values. Furthuremore, the characters such as punctuation, numbers, stop words, and URLs have been systematically eliminated. An important step of tokenization which is involves in data processing was executed using the deep learning and transformer models built-in tokenization functionalities and TF-IDF vectorization is done on the dataset for the SVM model.

In the pursuit of constructing an effective model for detecting fake reviews, a review of existing literature has been undertaken for the modelling of our amazon dataset. Numerous machine learning and deep learning models have been considered during this analysis. The LSTM and CNN models were employed due to their capabilities in capturing the temporal dependencies and extracting meaningful features from sequential dataset. For the process of classification task we have also taken into consideration the SVM classifier. We have also employed the tranfer learning models like BERT, RoBERTa, DistilBERT, and ALBERT. These models are already trained on vast datasets which enables their capabilities in discerning patterns in textual data which aligns with our motive in the Natural

Language Processing tasks. By combining these models together our approach toward detection of false reviews ensures our models model's adaptability to diverse patterns and contexts within the dataset.

4 Design Specification

The project architectural design of our study is illustrated in 1. A breakdown of each stage is presented in Section 5.

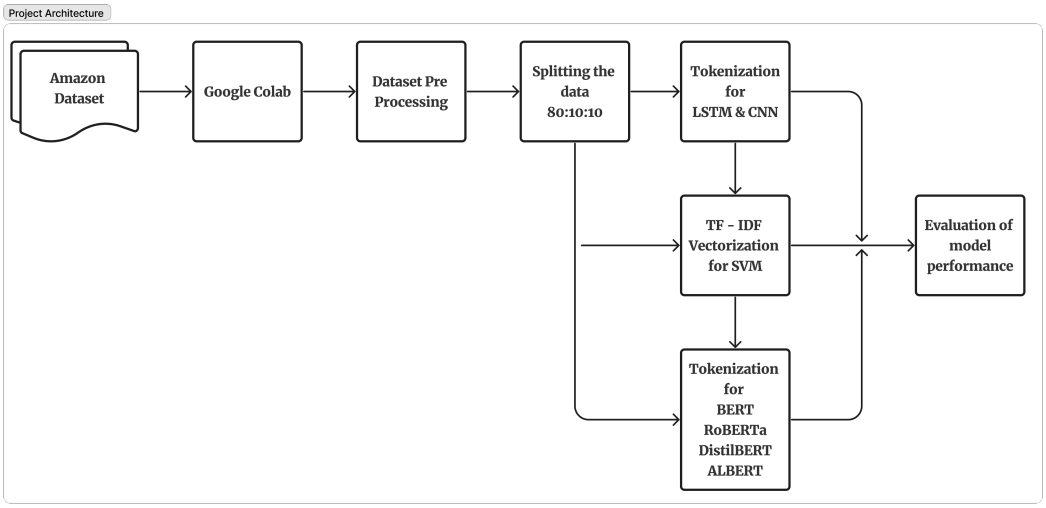

Figure 1: Model Architecture

4.1 Neural Network and Classifier

The LSTM which is a variant of RNN was designed to resolve the mitigating problem in RNNs. It utilizes a gating mechanism with memory cells which allows them to capture and retain information over an extended sequence. In our project it is beneficial because of the temporal dependencies inherited in the textual data which helps in identification of a pattern. CNN on the other hand which was designed for image analysis has the ability to employ convolutional layers to automatically extract hierarchical features from input data. In relation to our project it utilizes the filters to convolve over sequence which captures local patterns. Convolutional Neural Network (CNN), originally designed for image analysis, employs convolutional layers to automatically extract hierarchical features from input data. When applied to text, CNNs utilize filters to convolve over sequences, capturing local patterns and higher-level representations. In fake review detection, CNNs excel at discerning relevant linguistic features, making them instrumental in unveiling deceptive structures and linguistic nuances within reviews. The SVM models operates by constructing a hyperplane in a high-dimensional space to effectively classify data points. In the textual based context it utilizes the descision boundary to separate genuine and false reviews based on the extraction of features. To pull distinct features from the text, there must be conversion of text into numerical vectors for which standard vectorization is used such as Count Vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF). The word frequency is figured by counting the occurrence of each word in a sample

and then it is divided by the total numbers of words in the sample. On the other hand IDF is computed by taking the logarithm of the total dataset's number of documents and dividing it by the number of instances where a specific term is present.

4.2 Transformers

The design architecture of BERT (Bidirectional Encoder Representation from Transformer) lays the foundation for its remarkable capabilities in natural language processing tasks. BERT employs a transformer-based architecture, introduced by Vaswani et al. (2017), which revolutionized sequence-to-sequence learning by dispensing with recurrent neural networks. This architecture comprises encoder and decoder blocks, where BERT specifically focuses on the encoder. The design architecture of BERT employs a employs a transformer-based architecture, which made a ground breaking discovery in sequence to sequence learning with the help of RNN. The architecture consists of encoder and decoder blocks wherein it focuses on encoder more. Due to the multi-head self attention mechanism it is able to do bi-directional processing of input sequence which is essential for false review detection. Positional encoding is applied to encode the relative position of tokens which are in a sequence. It consists of 12 transformer block, which has 12 attention head and a hidden size of 768, resulting in 110 million parameters.

RoBERTa (Robustly optimized BERT approach), developed by Liu et al. (2019), builds upon BERT's architecture with refinements in training methodology. It eliminates the next sentence prediction task and incorporates a dynamic mask during training process. It was trained on the dataset for CC News and English Wikipedia which is 160GB in total. Roberta has 12 tranformer layers, 12 attention heads and and 768 hidden layers, amounting to 125 million parameters.

DistilBERT, a distilled version of BERT developed by Sanh et al. (2019), addresses BERT's computational inefficiencies while retaining its performance. It is 40% smaller and 60% faster which results in 97% of original model proficiency. DistilBERT-base-uncased, used in this research, possesses 6 transformers, each housing 12 self-attention layers, 768 hidden layers, and 66 million parameters.

ALBERT (A Lite BERT), introduced by Lan et al. (2019), optimizes BERT's architecture to resolve the training time and memory limitations. The model incorporates two-parameter reduction techniques, reducing GPU/TPU memory usage and accelerating training without compromising performance. By replacing the BERT's next sentence prediction with sentence order prediction (SOP) the ALBERT model overcomes NSP losses. The albert-base-v1 model, a pertained version, features 12 repeating layers, 12 attention heads, 768 hidden layers, 128 embeddings, and 11 million parameters.

5 Implementation

This section provides an in-depth explanation of the implementation process undertaken for the development of the fake review detection model in this research project.

5.1 Setup

The following table provides information regarding the programming languages, technologies, and libraries employed throughout the project.

Specification	Value
IDE	Google Colab
Computation	GPU
Number of GPU	1
Programming language	Python
Modeling library	SimpleTransformer, HuggingFace Transformer, Sklearn, Pandas, Numpy, Matplotlib, Seaborn, Wandb, Keras, PyTorch, tqdm, SciPy

Table 1: System Specifications for the Project

5.2 Data Loading and Pre-processing

In the first stages of Implementation we load the dataset. The CSV file of the dataset is first stored in the Google drive and then it mounted to the Google Colab Laboratory. The CSV file is then loaded to the a data frame using the pandas library. Our data was already cleaned and labelled because that was the requirement of the project hence there was no cleaning required. The we do the exploratory data analaysis. The data consists of 50% genuine and 50% fake reviews. There are 5 columns in the dataset out of which only the "Reviews" and "Fake_Reveiw_Flag" colums were picked for the development. The latter column is categorized in '1' and '0' which differentiates the reviews as genuine and fake.

5.3 Data sampling and Splitting

The data set consisted of 5000 records and to create an effective ML model the training is to be done rigorous, due to computational resource constrain using the sample function the data was trained at 10% and 50% and once the model was fined tuned as per the requirements it was trained at 100%. The sklearn library was used to split the data. It is divided in the Train-Validate-Test set using the using train test split function. The dataset was divide as follows 80% for the training, 20% for evaluation and test purpose.

5.4 Model Implementation

The training and implementation of the models are described in the sections below.

5.4.1 LSTM, CNN and SVM Models Implementation

To input the dataset into the model of LSTM and CNN, first we need to perform tokenization and sequence padding. The 'Tokenizer' class from the Keras library is employed to convert the raw text data into sequences of numerical tokens. The parameter num_words specifies the maximum number of words which is to be considered as feature in the tokenization process. This step helps in building a vocab of relevant words for the models. Next padding is done using the pad_sequence function which ensures that all the sequence have same length. The maximum length of words is given as 100, which means any review which has length longer that 100 words will be truncated and the shoter ones will be padded with 0. Then the tokenized data is input to the models.

To build the model of LSTM we start by employing an Embedding layer as the initial input layer, which transforms the tokenized sequence into dense vectors with a dimension

of 128 for each word. A single LSTM layer of 64 units is added and a dropout mechanism to counter the overfitting. The sigmoid activation function is used to facilitate binary classification for detecting the fake and genuine reviews. Then it is compiled with adam optimizer, and employing binary crossentropy as loss function and accuracy for metric evaluation. We train the data over multiple epochs which is validated against the test data.

For building the CNN model, similarly like LSTM we transform the tokenized data into vectors. Additionally we add 1D Covolutional layer and rectified linear unit which is responsible for extraction of features. The GlobalMAxPooling layer is responsible to select the most salient features. Binary classification and model compiling is done same as the LSTM model and then the evaluation is done using the accuracy metric.

To implement the SVM model we first use the TF-IDF vectorization technique. The TfidfVectorizer to convert the raw text reviews into a numerical format by calculating the TF-IDF weights, capturing the importance of each term in relation to the entire dataset. The resulting TF-IDF matrices, denoted as X_train_tfidf and X_test_tfidf for the training and test sets respectively which serves as the input features for training the SVM model. The SVM is constructed using a linear kernel which is the best choice for text classification. The training phase involves fitting the model to the TF-IDF training data. The performance is evaluated using the accuracy metric.

5.4.2 BERT, RoBERTa, DistilBERT and ALBERT Models Implementation

The four different transformer-based models follow a similar structure which encompases the tokenization, training and evaluation steps. The tokenization function which is specific to each model encodes the input data into tokenized sequence and attention mask. Once the datasets are tokenized models are loaded and moved to the GPU and are optimized using the adamw optimizer with a learning rate of 2e-5. The models were first run till 5 epochs but as the accuracy reached saturation, the epochs were reduced to 4. Then the test data is tokenized and the classification report prints the metrics. The key difference in all the models are as follows: BERT – it offers bidirectional context, RoBERTa – enhanced pre-training techniques, DistilBERT – provides a distilled version of BERT and ALBERT – it introduces parameter sharing mechanism for imporved performance.

6 Evaluation

This section discusses the performance of all the models which we have employed on the basis of metrics described below:

Accuracy: The percentage of correctly classified reviews.

Precision: The proportion of correctly identified false reviews among all predicted false reviews.

Recall: The proportion of correctly identified false reviews among all actual false reviews.

F1-Score: The F1 score, being the harmonic mean of precision and recall, provides a balanced assessment of a model's performance. A higher F1 score indicates better model performance, as it takes into account both precision and recall, contributing to a more comprehensive evaluation.

6.1 LSTM Model Evaluation

The LSTM model was trained for 5 epochs and the evaluation is as follows:

The training loss and accuracy in the first epoch were 0.2424 and 0.8986, respectively. The model's accuracy increased and its training loss decreased during the course of the next epochs, reaching a final epoch with an accuracy of 0.9842 and a training loss of 0.0454. Alongside training, validation metrics were tracked to provide light on the model's generalisation capabilities. Starting at 0.1821 and declining until the last epoch, the validation loss was 0.1907. In line with this, the validation accuracy steadily increased, peaking at 0.9435 during the most recent period. All of these findings point to the model's strong learning from the training set as well as its strong generalisation to new data. Standard classification measures were used to evaluate the LSTM model after training. The precision, recall, and F1-score for each of the two classes are reported in the classification report (0 and 1). Finding the right balance between recall and precision is essential in the binary classification job, because class imbalance can have serious consequences. An additional dataset evaluation showed an overall accuracy of 94%, demonstrating the model's ability to accurately classify cases.

Epoch	**Training Metrics**	**Validation Metrics**
1	loss: 0.2424, acc: 0.8986	val_loss: 0.1821, val_acc: 0.9327
2	loss: 0.1213, acc: 0.9537	val_loss: 0.1540, val_acc: 0.9446
3	loss: 0.0768, acc: 0.9730	val_loss: 0.1735, val_acc: 0.9350
4	loss: 0.0553, acc: 0.9804	val_loss: 0.1787, val_acc: 0.9452
5	loss: 0.0454, acc: 0.9842	val_loss: 0.1907, val_acc: 0.9435

Table 2: LSTM Training Summary

Examining class-specific measures in more detail, class 1 showed 0.94 precision, 0.95 recall, and 0.94 F1-score. When taken as a whole, these metrics highlight how reliable the model is at properly identifying examples. The model's balanced performance across classes is further highlighted by the weighted average and macro F1-scores, both of which are 0.94. In conclusion, after training over five epochs, the LSTM model demonstrated an impressive capacity for learning and generalisation, attaining balanced metrics and high accuracy in a binary classification test. The thorough evaluation metrics demonstrate the model's ability to discriminate between the two classes, indicating that it is a viable solution to the given classification problem.

	Precision	**Recall**	**F1-Score**	**Support**
Class 0	0.95	0.94	0.94	4118
Class 1	0.94	0.95	0.94	3969
Accuracy: 0.94 Macro Avg: 0.94, Weighted Avg: 0.94				

Table 3: LSTM Model Performance

6.2 CNN Model Evaluation

The CNN model was trained over 5 epoch. The training method produced an accuracy of 0.8814 and a loss of 0.2927 in the first epoch. The model showed notable improvement

as training went on, with the accuracy reaching a peak of 0.9993 by the fifth epoch and the loss dropping to 0.0053. This steady improvement in performance highlights how well the CNN can identify complex patterns in the training set. The model's performance was assessed on a different validation set at the end of each period. The measurements for accuracy and validation loss gave information about the model's capacity for generalisation. Throughout the training procedure, the CNN continuously showed good validation accuracy, reaching 0.9335. Interestingly, the validation loss did not increase significantly, suggesting that the model was not overfitting the training set.

Epoch	**Training Metrics**	**Validation Metrics**
1	33s/32ms - loss: 0.2927, acc: 0.8814	0.1956 - acc: 0.9293
2	9s/9ms - loss: 0.1099, acc: 0.9613	0.1765 - acc: 0.9362
3	7s/7ms - loss: 0.0441, acc: 0.9867	0.1902 - acc: 0.9363
4	9s/9ms - loss: 0.0154, acc: 0.9965	0.2258 - acc: 0.9336
5	7s/7ms - loss: 0.0053, acc: 0.9993	0.2446 - acc: 0.9335

Table 4: CNN Training Summary

The CNN was rigorously evaluated using conventional classification metrics on an independent dataset after training. For each class (0 and 1), the precision, recall, and F1-score were calculated. A well-balanced performance was found in the evaluation; class 0 obtained an F1-score of 0.93, a precision of 0.94, and a recall of 0.93. In a similar vein, class 1 showed 0.93 precision, 0.94 recall, and 0.93 F1-score. These findings demonstrate how well the model can discriminate between the two groups while striking a healthy balance between recall and precision. On the evaluation dataset, the CNN model's overall accuracy was 93%. Furthermore, weighted average and macro F1-scores, both at 0.93, highlighted the model's performance consistency across classes. This comprehensive analysis confirms that the CNN model performs well in binary classification and can effectively generalise to new data. In conclusion, the CNN model demonstrated strong learning throughout training, attaining high accuracy and minimal loss. The second assessment on a different dataset validated the model's proficiency in binary classification, as evidenced by balanced precision, recall, and F1-scores for both classes. These findings position the CNN model as a promising solution for the specific classification task it was designed to address.

	Precision	**Recall**	**F1-Score**	**Support**
Class 0	0.94	0.93	0.93	4118
Class 1	0.93	0.94	0.93	3969
Accuracy: 0.93 Macro Avg: 0.93, Weighted Avg: 0.93				

Table 5: CNN Model Performance

6.3 SVM Model Evaluation

The total accuracy of the Support Vector Machine (SVM) model is impressive, coming up at about 90.58%. The model's accuracy measure shows how well it can categorise examples into the appropriate classes. Precision, recall, and f1-score measures are used to further examine the model's performance, with a special emphasis on two classes: 0 and

1. The SVM model obtains a precision of 0.91 for Class 0, which includes 4118 instances, meaning that 91% of the positive cases that are predicted for this class are accurate. With a recall of 0.90 for Class 0, the model successfully accounts for 90% of the real positive cases. For Class 0, the f1-score—the harmonic mean of precision and recall—is 0.91. Taken as a whole, these metrics show how well the model can recognise examples that belong to Class 0. For Class 1, which has 3969 cases, the SVM model has a precision of 0.90, meaning that 90% of the positive cases that are predicted for this class are true. With a recall of 0.91 for Class 1, the model is able to accurately identify 91% of the real positive events. Class 1's f1-score is 0.90, indicating a performance that is balanced between recall and precision. The SVM model's accuracy of 0.91, when taken into account as a whole, indicates a strong capacity to accurately identify examples across both classes. The precision, recall, and f1-score macro-average and weighted-average metrics highlight the model's performance consistency even more. This thorough analysis shows that the SVM model can be trusted to handle the provided dataset in an efficient and dependable manner, boosting its prediction power.

	Precision	**Recall**	**F1-Score**	**Support**
Class 0	0.91	0.90	0.91	4118
Class 1	0.90	0.91	0.90	3969
Accuracy: 0.91				
Macro Avg: 0.91, **Weighted Avg**: 0.91				

Table 6: SVM Model Performance

6.4 BERT Model Evaluation

The table 7 shows the performance comparison of BERT model. The model's precision for Class 0 was 0.99, meaning that out of the cases that were predicted to be Class 0, 99% were in fact true positives. With a recall of 0.95 for Class 0, the model was able to correctly identify 95% of the real examples that belonged to Class 0. For Class 0, the F1-score—which weighs recall and precision—is 0.97. 4118 is the support value, which indicates how many instances there are in Class 0. The model's precision for Class 1 was 0.95, meaning it was 95% accurate in predicting occurrences of Class 1. At 0.99, the recall for Class 1 is exceptionally high, indicating that 99% of the instances in Class 1 were successfully caught by the model. Additionally, 0.97 is given as the F1-score for Class 1, indicating a balance between recall and precision. 3969 is the support value for Class 1. The model's overall accuracy of 0.97 indicates that 97% of the total cases were correctly classified. The precision, recall, and F1-score values for the macro-average, which gives the average performance across classes without taking into account class imbalances, are 0.97. The precision, recall, and F1-score weighted averages, which account for class imbalances, likewise show values of 0.97. In conclusion, the BERT model performs admirably in this classification job, obtaining 97% overall accuracy as well as strong precision, recall, and F1-score values for both classes. All of these measures together show that the model has a strong capacity to correctly categorise instances into the designated classes.

Class	Precision	Recall	F1-score	Support
0	0.99	0.95	0.97	4118
1	0.95	0.99	0.97	3969
Accuracy			0.97	8087
Macro Avg	0.97	0.97	0.97	8087
Weighted Avg	0.97	0.97	0.97	

Table 7: BERT Model Performance

6.5 RoBERTa Model Evaluation

The table 8 shows the performance comparison of RoBERTa model. The model exhibits an impeccable precision of 1.00 for Class 0, signifying that every instance that was predicted as belonging to Class 0 was, in fact, accurate. The recall value of 0.94 indicates that the model accurately classified 94% of the true instances belonging to Class 0. The F1-score, a metric that evaluates the performance of Class 0 by balancing precision and recall, is 0.97. A grand total of 4118 instances of Class 0 were processed by the model. Analogously, the precision value for Class 1 is 0.94, which indicates that 94% of the instances classified as Class 1 were indeed accurate. The recall value of 1.00 for Class 1 indicates that the model accurately classified every instance of Class 1. Class 1 has an F1-score of 0.97, which is consistent with the outstanding performance observed in Class 0. A grand total of 3969 instances of Class 1 were processed by the model. The model demonstrates an overall accuracy of 97% in both classes, denoting the percentage of instances that were accurately classified out of the entire set. The consistent performance of the macro average, which is calculated by averaging the metrics for each class irrespective of class imbalance, is evidenced by its precision, recall, and F1-score of 0.97. In consideration of class imbalance, the weighted average produces precision, recall, and F1-score all of which are 0.97. Based on a comprehensive dataset consisting of 8087 instances, these averages comprise both Class 0 and Class 1. In summary, the Roberta model exhibits exceptional precision, recall, and F1-score metrics across all classes, culminating in a remarkable overall accuracy of 97%. The consistent robust performance, even when macro and weighted averages are taken into account, across individual classes, indicates that the model is capable of managing the binary classification task.

Class	Precision	Recall	F1-Score	Support
0	1.00	0.94	0.97	4118
1	0.94	1.00	0.97	3969
Accuracy			0.97	8087
Macro Avg	0.97	0.97	0.97	8087
Weighted Avg	0.97	0.97	0.97	8087

Table 8: Roberta Model Performance

6.6 DistilBERT Model Evaluation

The table 9 shows the performance comparison of DistilBERT model.The precision value of 0.99 is assigned to Class 0, signifying that 99 percent of the instances classified as Class 0 were, in fact, accurate predictions. The recall value of 0.97 indicates that the model accurately identified 97% of the true instances belonging to Class 0. For Class 0, the

F1-score, which is the harmonic mean of precision and recall, is documented as 0.98. The aggregate of these metrics indicates a considerable degree of precision in discerning instances that are members of Class 0. Class 0 is designated to be supported for a total of 4118 instances. In the same way, the precision value of 0.97 is indicated for Class 1, which signifies that 97% of the instances classified as Class 1 were indeed accurate. The recall value of 0.99 for Class 1 signifies that the model correctly classified 99 percent of the true instances belonging to that class. Additionally, the F1-score of 0.98 is provided for Class 1, indicating a well-balanced performance with respect to both recall and precision. Class 1 is specified to have a support of 3969 instances. As indicated by the aggregate accuracy value of 0.98 for the DistilBERT model, 98% of the instances in the dataset were classified accurately. The performance of both classes is deemed average by the macro-average, which yields precision, recall, and F1-score values of 0.98. The weighted average, which accounts for possible class imbalances, produces precision, recall, and F1-score values of 0.98 as well. The averaged metrics provide a thorough evaluation of the model's efficacy in managing both classes; the high values signify a robust overall performance on the 8087-item dataset that was assessed.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.99	0.97	0.98	4118
1	0.97	0.99	0.98	3969
Accuracy			0.98	8087
Macro Avg	0.98	0.98	0.98	8087
Weighted Avg	0.98	0.98	0.98	8087

Table 9: DistilBERT Model Performance

6.7 ALBERT Model Evaluation

The table 10 shows the performance comparison of ALBERT model. The model's capacity to categorise examples into the two classes, represented by the numbers 0 and 1, is assessed in this context. In addition to overall accuracy, macro-averaged scores, weighted-averaged scores, and precision, recall, and f1-score for each class, these assessment measures are also included. With a precision of 0.95 for Class 0, 95% of the instances that were predicted to be Class 0 are in fact true positives. With a recall of 0.99 for Class 0, the model is able to accurately identify 99% of the real cases that correspond to the class. The matching f1-score, which strikes a compromise between recall and precision, is 0.97. The real number of instances in Class 0 is 4118, which is the support for this class. Class 1 displays performance data that are marginally different. Class 1 precision is 0.99, indicating a high degree of accuracy in Class 1 instance prediction. On the other hand, Class 1 recall is 0.95, meaning that 95% of Class 1 occurrences are captured by the model. Class 1's f1-score is likewise 0.97, indicating a harmony between recall and precision. 3969 is the support number for Class 1. With an overall accuracy of 0.97 for both classes, the model is able to accurately predict the class labels for 97% of the dataset's cases. The unweighted average of recall, f1-score, and precision for both classes is provided by the macro-averaged scores, which come out to be 0.97. Similarly, values of 0.97 are also obtained using the weighted-averaged scores, which account for the class distribution. To summarise, the ALBERT model exhibits remarkable performance in binary

classification, attaining high recall, f1-scores, and precision for both classes, culminating in an astounding 97% overall accuracy.

Class	Precision	Recall	F1-Score	Support
0	0.95	0.99	0.97	4118
1	0.99	0.95	0.97	3969
Accuracy			0.97	8087
Macro Avg	0.97	0.97	0.97	8087
Weighted Avg	0.97	0.97	0.97	8087

Table 10: ALBERT Model Performance

7 Discussion

7.1 Performance Comparison of LSTM,CNN & SVM

The graph in 2 incorporates both bar and line components, creating a combination chart that aims to compare three distinct machine learning models: LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), and SVM (Support Vector Machine).

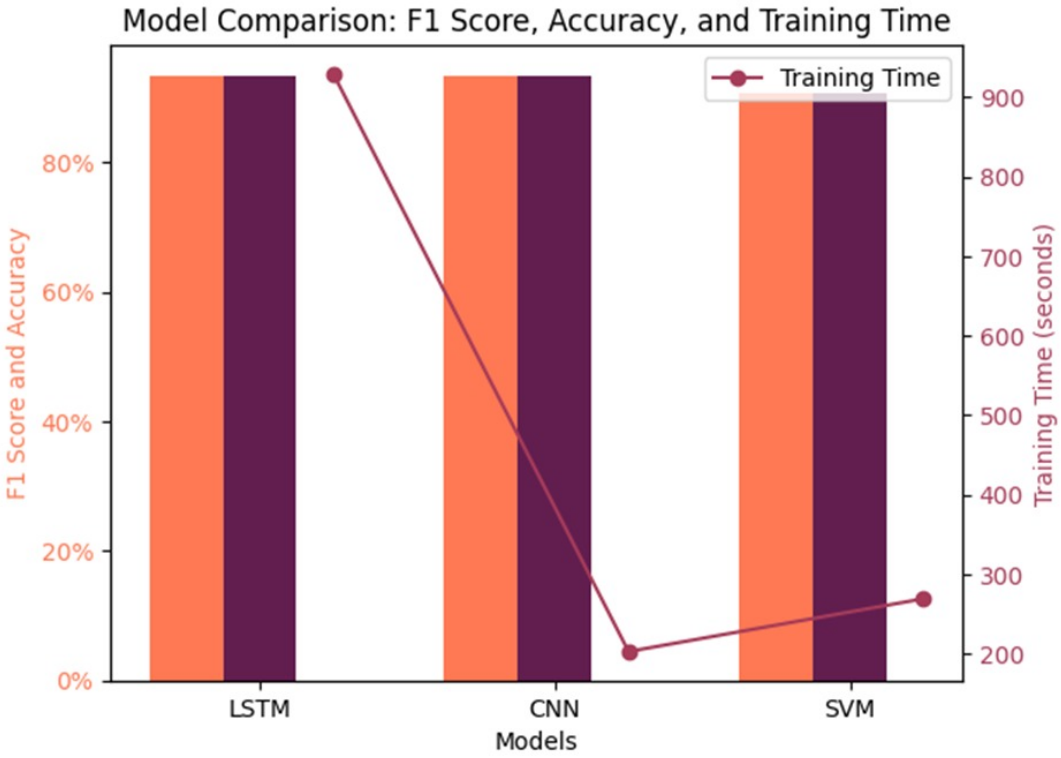

Figure 2: Model Architecture

The comparison revolves around two crucial performance parameters, namely the F1 Score and Accuracy, which are visualised through a bar chart. Additionally, the line chart represents the Training Time.

In the bar chart depicting the F1 Score and Accuracy, it is observed that both metrics demonstrate commendable values across all models, surpassing the established threshold

of 60% for each. LSTM has an accuracy of 94% and CNN has an accuracy of 93% for which the F1 Score is 0.94 and 0.93 respectively , while the SVM demonstrates slightly lower metrics for both criteria with and accuracy of 91% and F1-score of 0.90. The F1 Scores and Accuracy of these models have consistently remained at a high level, indicating that the training has been successfully done. With the accuracy and F1-score it is evident that the LSTM has outperformed but taking in consideration the computational time we can see that CNN has the shortest time period with lower than 100s of evaluation and the SVM model follows with 258 seconds in evaluation wherein LSTM computes in over 2000s.

LSTM performs well, but its lengthy training duration may make it unsuitable for quick model training. CNN, on the other hand, finds a good balance between results and training time, making it a good choice for detection of reveiw. Notably, SVM stands out because it has a short training time, which makes it a good choice for situations where time or computing power are limited, even though its F1 Score and Accuracy measures are slightly lower than the other two models.

7.2 Performance Comparison of BERT, RoBERTa, DistilBERT & ALBERT

This bar chart in 2 with a line chart compares four different transformer-based machine learning models: BERT, RoBERTa, DistilBERT, and ALBERT.

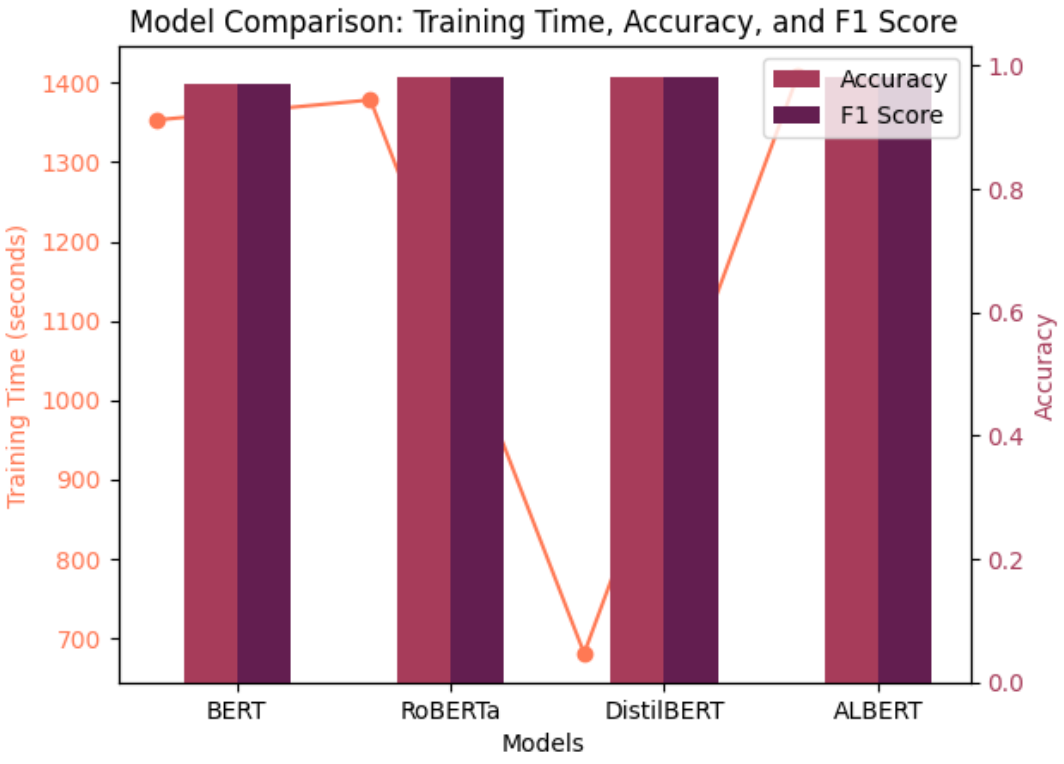

Figure 3: Model Architecture

The bar chart shows the training time, and the line chart shows the accuracy and F1 score. The line graph indicates that the training durations of BERT and RoBERTa which is 1350 seconds and 1380 seconds respectively are comparable, as both demon-

strate the most extensive training periods out of the four models evaluated. DistilBERT distinguishes itself through its significantly reduced training time of about 680 seconds, which is in accordance with its intended design goals of producing a model that is both lighter and quicker than the other models. The training period of ALBERT is situated between that of DistilBERT, BERT and RoBERTa. The bar chart, which probably uses the secondary y-axis on the right side to measure, shows that all models have consistently high accuracy above the threshold of 60% and F1 scores above the cutoff of 0.8, Impressively, DistilBERT's performance is the best compared to other models, with an accuracy of 98% and F1 score of 0.98. This is especially impressive given how little time it spent training compared to BERT and RoBERTa.

From the model we can infer that the DistilBERT is the best model out of all. In line with our goal of producing an efficient model in detection of reviews, DistilBERT proves to be the most efficient.

7.3 Performance Comparison of All Models

The above bar chart shows a comparison of the accuracy metrics of different ML models for the purpose of bot or automated review detection:

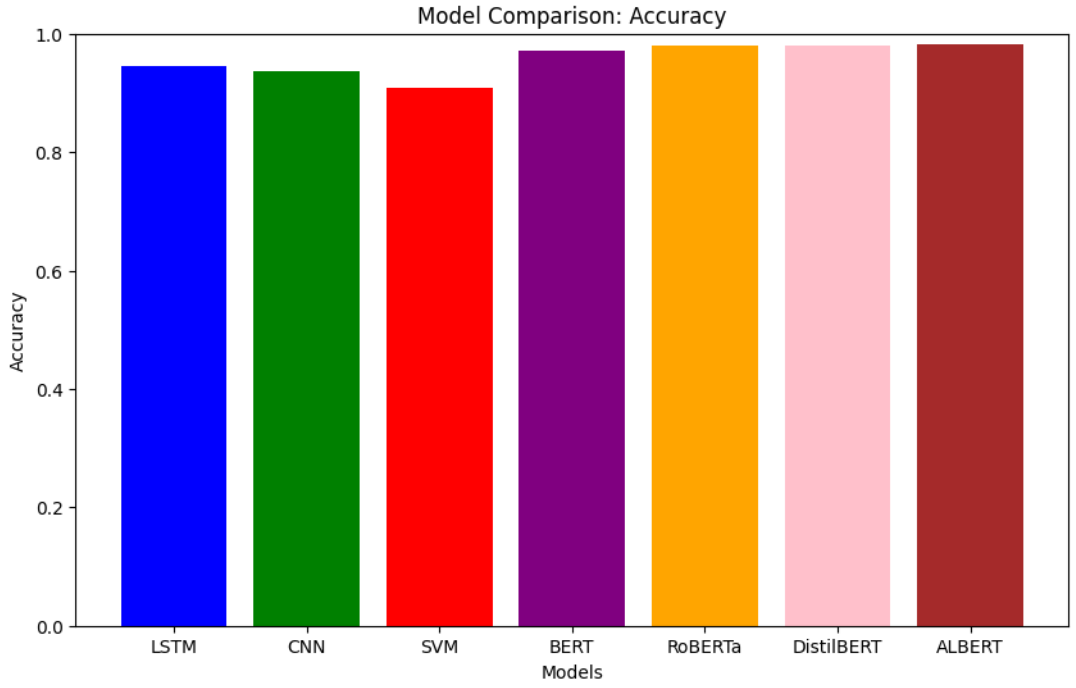

Figure 4: Model Architecture

Sequential processing (LSTM) and spatial hierarchy (CNN) are compared in the study of performance insights, and their nearly identical correctness indicates a remarkable similarity in their efficacy for the assessed task. Remarkably, the accuracy attained by the non-neural network model Support Vector Machine (SVM) suggests that, in certain situations, it might be a good substitute in the event of false detection task. Additionally, a comparison of other transformer models, such as BERT, RoBERTa, DistilBERT, and ALBERT, highlights their continuously high accuracy, indicating a high degree of effectiveness in the context of the assessed task, which is probably related to natural language

processing (NLP). The superiority of DistilBERT over BERT that has been noticed is consistent with the improved architecture of the former.

DistilBERT's presentation is noteworthy since it keeps excellent accuracy while perhaps providing benefits in terms of inference speed and model size. This result emphasises the advantages and possible trade-offs of various transformer topologies. Although ALBERT's accuracy is marginally less than DistilBERT's, its design choices might make it a better option when memory footprint and model size restrictions are critical. To sum up, DistilBERT's proves to be the best model in comparison to all the models and will be efficient choice for the task our false review detection.

7.4 Performance Comparison to Previous Research

In the research, on the basis of result we can see that the CNN model was noted for its efficiency which had completed its evaluation in 100 seconds which made the model an adequate choice for the task of detection while the SVM model which was still slower than the CNN model with an evaluation time of 258 seconds can be considered efficient in the scenarios where there is limitation in computing resources and lastly the LSTM model despite its high performance had a significant higher computation time of over 2000 seconds which is a drawback for our research project but in comparison with other research we can see that (Gautam et. Al. 2022) discussed various ML and deep learning algorithms which included SVM and LSTM. Notably the Bi-LSTM model which an enhanced version of LSTM was found to be best performing among all the LSTM variants achieving an accuracy of 93.75% at epoch 19 while the LSTM model was able to achieve and accuracy of 85%, whereas in our research the LSTM model was able to achieve an accuracy of 98.42% at epoch 5. Furthermore, results for SVM model we can see that (Noori et. Al, 2021) was able to achieve an accuracy of 68.58% whereas our research was able to achieve an accuracy of 91%.

For the evaluation of transformer-based models, from the research work in (Vinayagamurthy et al, 2022) the BERT model with an epoch setting of 4 showed an accuracy of 93%, also in another research by (Gupta et al.; 2021) the performance of BERT, RoBERTa, DistilBERT, and ALBERT was reported with accuracies of 65%, 67% 65% and 64% respectively, whereas in comparison with our research with an epoch of 4, the models BERT, RoBERTa, DistilBERT, and ALBER was able to achieve an accuracy of 97% , 97%, 98% and 97% respectively. We can clearly say that our models demonstrated superior performance in terms of accuracy indicating their robustness in handling complex tasks of review detection

8 Conclusion and Future Work

The reviews found on e-commerce platforms play a vital role in shaping the descision of a customer and can have a direct impact on the financial outcomes of a business. When counterfeit reviews are crafted with a motive to mislead a consumer hence it becomes important to accuractely identify bot based reviews. There has been significant research in this domain but the challenges always persist such as the necessity of the domain expertise to conduct well verse feature engineering in ML projects which also requires a labelled datasets in the ear of deep learning. In this study we investigated the effectiveness of multiple ML models for the task of bot or automated text detection. We compared sequential processing of LSTM, spatial hierarchy of CNN and non-neural network SVM. Moreover also the transfer based models like BERT, RoBERTa, DistilBERT, and ALBERT were evaluated. These models undergo training using Amazon dataset, aiming to create a robust bot review classifier while minimizing computational resource usage. The process in training the model involved data pre-processing, hyperparameter selection and fine-tuning of the transformer models. The performance of all models are evaluated based on the performance metrics of accuracy, weighted F1-score and computational time. The results showcased that the LSTM, CNN and SVM proving to be a reliable alternative out of which CNN out performed the other two with an accuracy of 93% and F1-score of 0.93 and computational time of 100s. On the other hand, transfer models exhibited high accuracy, emphasizing their efficiency in NLP tasks. Keeping the performance metrics in mind notably DistilBERT model proves to be the standout performer across BERT, RoBERTa and ALBERT, with an accuracy and F1-score of 98% and 0.98 respectively and merely spending 680 secs in the computation of the model, making it apparent as a choice for bot review detection in our study.

As we conclude the project, future avenues can involve exploration of ensemble models, combining different models or employing ChatGPT, GPT-2, and XLNet by OpenAI or Sparrow, LaMDA, PaLM, and T5 are developed by Google AI, which could prove to enhance the performance.

Acknowledgement

I would like to express my sincere gratitude to Professor Aloka Anant for his constant support, invaluable guidance, and continuous encouragement throughout the course of my project. His expertise, insightful feedback, and commitment to the academic growth have been instrumental in shaping the quality and direction of my work. I would also like to thank the National College of Ireland for providing me with the opportunity to undertake this project. The academic environment, resources, and infrastructure offered by the institution have significantly contributed to the successful completion of this research. I deeply appreciate for the support I have received from both my professor and the National College of Ireland, which has played a crucial role in my academic and professional development. Thank you.

References

Abdul, S., Qiang, Y., Basit, S. and Ahmad, W. (2019). Using bert for checking the polarity of movie reviews, *International Journal of Computer Applications* **177**.

Desai, N., Nandinini, K. and Jahnavi, P. (2023). Fake review detection using deep learning, *JETIR* **10**: Not specified. April 2023.

Deshai, N. and Rao, B. B. (2023). Unmasking deception: A cnn and adaptive pso approach to detecting fake online reviews, *Soft Computing* **27**: 11357–11378.
URL: *https://doi.org/10.1007/s00500-023-08507-z*

Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding, *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Frontiers (2021). A study on the influence of online reviews of new products on consumers' purchase decisions: An empirical study on jd.com, `https://www.frontiersin.org/articles/10.3389/fpsyg.2021.631366/full`. Accessed: [date].

Gupta, P., Gandhi, S. and Chakravarthi, B. R. (2021). Leveraging transfer learning techniques- bert, roberta, albert and distilbert for fake review detection. in ACM International Conference Proceeding Series.

Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X. and Li, G. (2016). Deceptive review spam detection via exploiting task relatedness and unlabeled data.

Hassan, R. and Islam, M. (2020). A supervised machine learning approach to detect fake online reviews, *ICCIT 2020 - 23rd International Conference on Computer and Information Technology*.

He, P., Liu, X., Gao, J. and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Comput.* **9**.

Hu, Y., Ding, J., Dou, Z. and Chang, H. (2022). Short-text classification detector: A bert-based mental approach, *Comput Intell Neurosci* .

Jindal, N. and Liu, B. (2007). Review spam detection, *WWW '07 Proceedings of the 16th international conference on World Wide*, pp. 1189–1190.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis, *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 219–230.

K., S., D., M., S., W., D., L., H., C. and H., L. (2019). User-centric fake news detection: A comparative study of machine learning and deep learning approaches.

Khan, H., Asghar, M., Asghar, M., Srivastava, G., Maddikunta, P. and Gadekallu, T. (2021). Fake review classification using supervised machine learning, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Kim, J., Kang, J., Shin, S. and Myaeng, S. (2021). Can you distinguish truthful from fake reviews? user analysis and assistance tool for fake review detection. Link.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* .

Mir, A. Q., Khan, F. Y. and Chishti, M. A. (2023). Online fake review detection using supervised machine learning and bert model, *Not specified* **Not specified**: Not specified. Not specified.

Ott, M. (2018). Deceptive opinion spam corpus v1.4. [Online]. `http://myleott.com/op-spam.html`.

Ott, M., Cardie, C. and Hancock, J. T. (2013). Negative deceptive opinion spam, *Proceedings of NAACL-HLT 2013*, Atlanta, Georgia, p. 497–501.

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

Shan, G., Zhou, L. and Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection, *Decision Support Systems* **144**: 113513.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**: 1929–1958.

Taşağal, K. and Uçar, (2018). Detection of fake user reviews with deep learning, *International Journal of Research in Engineering and Applied Sciences (IJREAS)* **8**(12). Available at SSRN: `https://ssrn.com/abstract=3319640`.

Tellawar, A., Wattamwar, S. and Thorat, B. (2023). Online fake review detection using svm, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* **11**: Not specified. Available at www.ijraset.com.
URL: *www.ijraset.com*

Varlamis, I., Nasir, J. A. and Khan, O. S. (2021). Fake news detection: A hybrid cnn-rnn based deep learning approach, *International Journal of Information Management and Data Insights* . In Press.

Vyas, P., Liu, J. and El-Gayar, O. (2021). Fake news detection on the web: An lstm-based approach. Faculty Research & Publications, 267.
URL: *https://scholar.dsu.edu/bispapers/267*

Wang, C.-C., Day, M.-Y., Chen, C.-C. and Liou, J.-W. (2018). Detecting spamming reviews using long short-term memory recurrent neural network framework, *Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government* pp. 16–20.