An Integrated Approach: Fake Review Detection using convBERT-BiLSTM Classification

by

Md. Anas Mahmud
20101149

Alina Hasan
20101301

Tajrian Mahbub
20101325

Navid Hasan Rafi
20101585

Rushayed Ali Faiaz
21301717

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2024

© 2024. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Anas Mahmud
20101149

Alina Hasan
20101301

Tajrian Mahbub
20101325

Navid Hasan Rafi
20101585

Rushayed Ali Faiaz
21301717

Approval

The thesis/project titled "An Integrated Approach: Fake Review Detection using convBERT-BiLSTM Classification" submitted by

1. Md. Anas Mahmud (20101149)
2. Alina Hasan (20101301)
3. Tajrian Mahbub (20101325)
4. Navid Hasan Rafi (20101585)
5. Rushayed Ali Faiaz (21301717)

As of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 22, 2024.

Examining Committee:

Supervisor:
(Member)

____

Md. Khalilur Rahman, phD
Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

____

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

____

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In the era of E-commerce, online reviews significantly shape consumer buying decisions and store evaluations. However, the prevalence of unethical practices such as review manipulation poses a considerable challenge. Businesses often hire spam reviewers or deploy bots to boost their reputation or even damage that of their competitors. Despite existing efforts in the field of fake review detection, there remains a need for further studies. In contribution, we propose the development of a scoring rubric designed to guide annotators in the identification of fake reviews and a hybrid model ConvBERT-BiLSTM for detection. We leverage the efficiency of ConvBERT, a compact variant of the BERT model, and the superior capabilities of BiLSTM over LSTM. The model is trained on a dataset gathered from Amazon. The dataset comprises 7,727 labeled reviews using the rubric. Through careful assessment, the proposed model garnered an accuracy of 97% surpassing previously established BERT variants.

Keywords: Natural Language Processing, Fake Review Detection, Human Method, Transformers, Neural Networks, BERT, ConvBERT, LSTM, BiLSTM

Acknowledgement

First and foremost, all praise to Allah Subhanahu Wa Ta'ala under whose blessings our thesis has been successfully concluded without any major disruptions. Second of all, we are deeply grateful to our supervisor Dr. Md. Khalilur Rahman for this opportunity. This thesis would have not been completed without his guidance and support throughout our research.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ALBERT A Lite BERT

BERT Bidirectional Encoder Representations from Transformers

BFRD Bengali Fake Review Dataset

BiLSTM Bidirectional Long Short-Term Memory

CNN Convolutional Neural Network

ConvBERT Convolutional BERT

DBI Davies–Bouldin index

ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately

FN False Negative

FP False Positive

GAN Generative Adversarial Network

IQR Interquartile Range Filtering

KNN K-Nearest Neighbors Algorithm

LSTM Long Short-Term Memory

MLM Masked Language Mode

NLP Natural Language Processing

NSP Next Sentence Prediction

RNN Recurrent Neural Network

RoBERTa Robustly Optimized BERT Approach

SC Silhouette Coefficient

SEM Semantic and Emotion Model

SVM Support Vector Machine

$T5$ Text-to-Text Transfer Transformer

$TF-IDF$ Term Frequency-Inverse Document Frequency

TN True Negative

TP True Positive

Chapter 1

Introduction

1.1 Background

What are fake reviews? Fake reviews are deceptive reviews that are written to manipulate the consumers' buying decisions. This can be created either by individuals or automated bots and is often used to promote or slander goods or services offered by a business. Some of these businesses engage in these unethical practices to gain an unfair competitive advantage against their rivals. These businesses create positive fake reviews to promote their reputation and can conversely, create negative reviews to damage the reputation of their competitors.

The COVID-19 pandemic has revolutionized E-commerce, making online shopping an integral part of modern society. In 2020 alone, E-commerce sales witnessed a remarkable 43% increase compared to pre-pandemic years [27]. The influence of online reviews is undeniable, as shown in research conducted by Dixa, revealing that a staggering 93% of consumers read reviews before making a purchase [38]. However, this has brought forth a concerning issue: the prevalence of fake reviews. As customers become more reliant on online reviews, the impact of fake reviews becomes more alarming. A survey conducted by BAZAAR Voice found that nearly 97% of respondents lose trust in a brand once they spot a fake review [33].

As stated by Chen et al (2017) online reviews can help potential consumers make purchase decisions, but few leave feedback after completing the deal [7]. Sellers often commit trust fraud to boost their reputation and attract more consumers, such as employing professional scammers or offering rebates to generate positive reviews. Consumers rarely submit negative reviews because they fear sellers might repeatedly pester them by telephone to soften a negative review. Wang et al (2018) state that on the other hand, people's reviews depend on their taste and mentality, he gives an example of washing machine noises as the noise can be high for some people but the true color of the noise is not that high [11]. So the reviews from the users could not be correct for everyone.

Jindal and Liu classify fake reviews into three categories: untrue, brand-focused, and irrelevant [36]. Many platforms have adopted strategies to combat these types of reviews. For example, Amazon has a dedicated team of experts who detect and block fake reviews, and in 2020, they reported removing 200 million suspected reviews [25]. However, despite such efforts, the challenge persists as Fakespot reports that approximately 42% of the 720 million reviews on Amazon require revision, indicating a significant number of undetected fake reviews that even the platform itself may miss [35].

1.2 Problem Statement

Nowadays, Corporations are becoming more strategic. Recognizing the potential of positive reviews, these businesses generate favorable ratings in mass to manipulate the review system. Hence, the average rating becomes overwhelmingly positive and sways customers into buying their goods and services. Likewise, some businesses also generate fake negative reviews on competitor platforms. The following passage will reveal three common yet popular methods for creating such reviews.

The first method involves dedicated pages or groups from which positive reviews can be bought. By purchasing these so-called "Boosting" packages, businesses hire the assistance of providers who generate reviews using either bots or authentic accounts [31]. UseViral and SidesMedia are some examples of websites that facilitate fake review boosting. Similarly, bad reviews, downvotes, or dislikes can be bought to damage the reputation of their competitors [12].

The second method entails the recruitment of individuals who purchase their products and subsequently leave a 5-star review labeled "Authentic Purchase" or "Verified Purchase" [18]. They are then reimbursed by the company, often with an additional commission ranging between $5 and $10.

The third method is rather tricky. Often sellers exploit platforms that allow them to edit product details while retaining previous positive reviews. This is a practice particularly observed on Amazon and Daraz, where sellers initially sell a good product and accumulate positive reviews, then modify the product listing to display a completely different item, usually of subpar quality or unrelated to the original category. For instance, a nice pair of socks may be replaced with a defective electronic. The unchanged positive reviews mislead customers into believing that the fraudulent item is "Most Popular" [9].

1.3 Research Objectives

Our research aims to contribute to the field of fake review detection using Natural Language Processing (NLP) to extract features from the textual data in reviews. Additionally, using feature engineering, the extracted features may be further processed to either create more meaningful features or, better capture any underlying patterns in the data. Finally, we will need to develop a model that not only preserves contextual data but is capable of understanding it.
For this purpose, our objectives are to

1. Gather reviews from a prominent online shopping platform for annotation.
2. Investigate techniques related to NLP.
3. Investigate techniques related to feature engineering.
4. Identify models that align with our criteria.
5. Develop a novel architecture that effectively classifies reviews.
6. Compare the performance of the proposed model with other classifiers.
7. Disseminate our research findings via conference presentations and journal papers.

1.4 Document Outline

In the upcoming chapters, we delve into the challenges of fake review detection, explore existing studies and methodologies, and introduce a novel model aimed at fostering fair competition and bolstering the credibility of online shopping platforms. Chapter 2 provides an in-depth literature review of established models in deception research and related fields. Following that, Chapter 3 outlines the overall workflow, dataset, and our proposed model. In Chapter 4, we present and compare the experimental results of the models. Finally, Chapter 5 concludes our study.

Chapter 2

Literature Review

Following the advent of E-commerce, there is growing concern over whether a human or bot generated a review. Researchers and practitioners in the industry have created a variety of algorithms based on natural language processing (NLP) to identify fake reviews. The passage delves into a collection of papers that have investigated various techniques to gain a better comprehension of the matter at hand.

2.1 Related Work

Feature extraction plays a crucial role in NLP and therefore various factors require examination. In one study, reviewer behavior, helpfulness, comments, and search queries were analyzed to determine the authenticity of reviews collected from Epinions [3]. Similarly, another study considered the statistical analysis of diverse behavioral features – activity window, review count, reviewer deviation – alongside linguistic n-grams [5]. The study confirmed a link between unusual reviewing patterns and fake reviews, with n-grams offering mild accuracy gains. The application of the Random Forests classifier on Yelp data, reveals that concentrating on review and reviewer-centric features provides a robust approach for identifying singleton fake reviews [8]. In a comparative analysis of fake review detection, the SVM model demonstrated the highest precision and F1 score, while KNN had the highest recall [13]. Random Forest achieved the highest accuracy, but it was noted that accuracy may not be the most suitable metric for evaluation. The study emphasized feature selection using TF-IDF. Moreover, feature engineering too plays a significant role, as portrayed in an article that employed logistic regression and achieved an accuracy of 88%, outperforming other algorithms [41]. In the same research, Yelp restaurant and hotel reviews, and approximately 40,000 Amazon reviews were analyzed using sentiment analysis. By applying sentiment thresholds and using Random Forest classification, a high accuracy of 91% was attained.

Bidirectional Encoder Representations from Transformers (BERT) has been a game changer in the field of NLP since 2018, achieving cutting-edge performance in a variety of NLP-based tasks. In a deception detection paper various models, including context-aware and context-free variations, were used to classify labeled utterances [21]. The contextualized BERT-based models equipped with a transformer surpassed the SVM classifier that produced an accuracy of 70% by 1-2%. This highlights the importance of linguistic context for improved detection. Additionally, BERT has also been applied in the field of fake review detection. Equipping the model, allows classifiers to better comprehend word relationships, resulting in a considerable increase in the accuracy: The SVM from 80% to 87%, Random Forest from 79% to 83%, and KNN from 71% to 77% [39].

The impact of BERT has extended to many distinctive versions. ELECTRA is used for pre-training Transformer text encoders and provides faster training and higher accuracy in downstream tasks than BERT [16]. Its pretraining includes a discriminator that detects replaced tokens, while a generator predicts the original tokens for hidden tokens in corrupted sequences. BanglaBERT based on the ELECTRA improves the text classification accuracy in Bengali [26]. In one of the approaches, BanglaBERT outperformed deep learning models, achieving the highest weighted F1 of 0.809 on the publicly available dataset “Bengali Fake Review Detection” dataset [40]. Previously recognized for its ability to outdo more state-of-the-art methods in fake review detection on a benchmark deception dataset with an accuracy of 91% [23], RoBERTa optimizes BERT by tuning hyperparameters and training scaling [15]. On the other hand, the XLNet combines auto-regressive and auto-encoding models to improve contextual understanding [20]. Both models have already been tried and tested on the “Kaggle Deceptive Opinion” dataset containing 1600 recordings [30] with each surpassing deep learning models with an accuracy of 94% and 97% respectively. Moreover, ALBERT also known as “A Lite BERT” is a self-supervised learning model that works on increasing speed and lowering memory consumption using a parameter reduction technique [17] while the DistilBERT transforms BERT features into a more efficient model for faster inference [19]. In a recent study, utilizing 50% of the large Yelp dataset comprising 1.4 million records labeled fake (0) or genuine (1), variants of BERT namely ALBERT and DistilBERT were trained to identify fake reviews [29]. ALBERT attained a weighted F1 score of 0.66 whereas DistilBERT gained a slightly higher score— 0.68. These variations demonstrate the versatility and adaptability of BERT.

Long Short-Term Memory Model (LSTM) is often employed as a deep learning classifier in the field of detection. One such study combines the model with the auto-encoder, an effective approach for unsupervised spam review detection [32]. The hybrid model preserves the local context of reviews and outperforms global context-based models with an accuracy of 0.34 DBI and an SC score of 0.71. Another study involves the CNN-LSTM architecture, which combines CNN layers for feature extraction and LSTM networks for sequence prediction to detect fake user comments [10]. It achieved an accuracy of 87% which is higher than that of the LSTM and BiLSTM models.

Some notable studies include use of SEM to detect fake reviews, with a focus on identifying unbalanced emotions [6]. Naive Bayes, SVM, and Decision Tree classifiers were used for this purpose. Then we have the development of ScoreGAN, a variation of GAN, which is designed to incorporate ratings, distinguish between bots and human-generated reviews, and enhance training stability [34].

Chapter 3

Methodology

3.1 Workflow

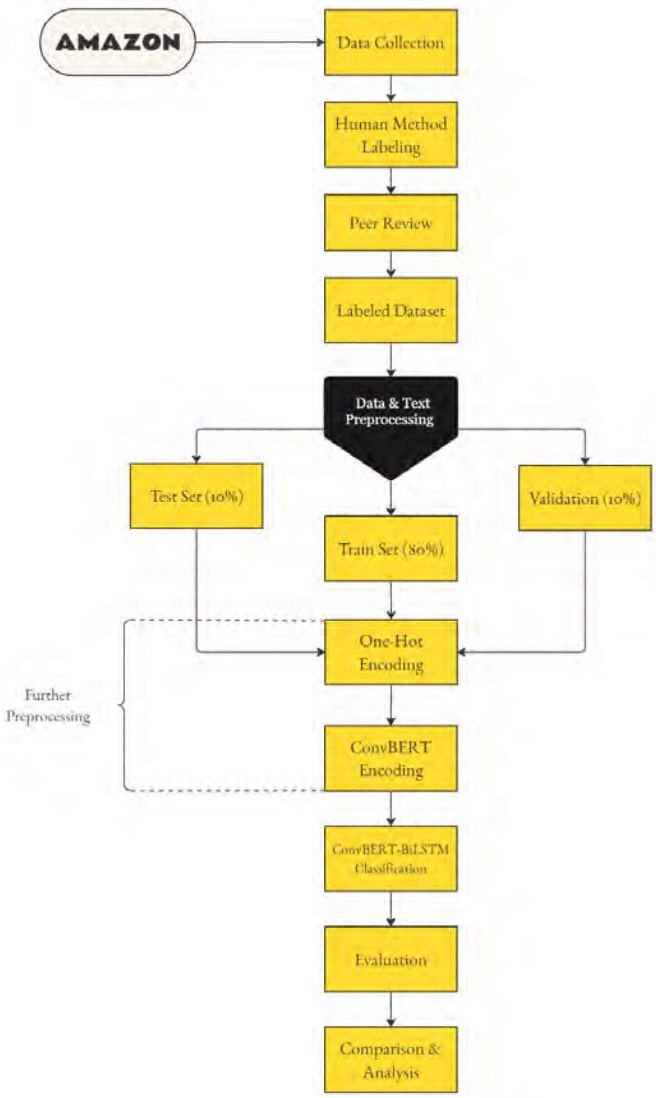


Figure 3.1: Workflow

With a multi-pronged strategy, we delved into the research of fake review detection. The first step involved developing a scoring rubric based on research studies and articles on identifying fake reviews to guide potential annotators; the subsequent steps included data collection from Amazon, annotating using the human method followed by a peer review. The labeled dataset was then preprocessed and partitioned into training, testing, and validation sets. Next, the labels undergo one-hot encoding. Our proposed model is a novel approach to fake review detection. It combines the functionalities of ConvBERT for feature extraction and two stacked BiLSTMs for classification. The ConvBERT encoder tokenizes the textual data. Finally, the hybrid model was trained on the training set, and assessed on the validation and testing sets. The outcome was evaluated using metrics, F1 score accuracy, and weighted average. A comparative examination was then conducted against previously established models. Throughout this study, we scrutinized the results, meticulously examined any errors, and studied the broader implications to further refine methodologies in fake review detection of online shopping platforms.

3.2 Dataset

3.2.1 Scoring Rubric

Historically, efforts to discern review spam have leaned on heuristic rules like helpfulness voting or evaluating rating deviations from the average product rating [3]. A more principled approach to the annotation of spam reviews trains ten college students on articles related to the identification of spam reviews [3]. Subsequently, they were tasked to label a dataset independently. Each review was then analyzed by two individuals, with discrepancies resolved by a third.

Similarly, we plan to circumvent this issue but instead of training students on analogous studies, a scoring rubric which is further solidified with the approval of present industry experts is developed instead. This rubric serves as a guideline for our annotators, enabling them to crosscheck and provide comments on specific aspects of the reviews. The intention is to foster a multi-perspective approach, ensuring a more consistent labeling process using the human method.

Initially, each reviewer is assigned a base score of 9. Points are then, either rewarded or deducted from the score in accordance to the rubric provided in Figure 3.2.

Feature	Type	Cost
Suspiciously Priced	yes but unacknowledged	-0.5
	yes and acknowledged	+0
	no	+0
Average Rating	greater than or equal to 4.5	-0.5
Username	peculiar, numbered & rarely derives meaning	-1
Date	multiple same date reviews	-1
Verified Purchase	yes	+0
	no	-1
Number of Helpful Votes	0	+0
	1 to 5	+0.2
	6 to 10	+0.4
	11 to 15	+0.7
	16 or more	+1
Image	attached	+0
	none	-0.5
Vine Review	yes	-1
	no	+0
Review Details	incoherent phrasing	-0.5
	excitable phrasing	-0.5
	somewhat matches product details	-0.5
	does not match product details	-9
	too short or detailed	-0.5
	repetitive usage of words within review or across reviews	-0.5
Review-Rating Polarity	completely positive review with good rating	-0.5
	completely negative review with poor rating	-0.5
	somewhat negative review with good rating and vice versa	-0.5
	completely negative review with good rating and vice versa	-9

Figure 3.2: Scoring Rubric

Suspiciously Priced: Certain businesses may offer products at prices that may be too good to be true. Thus, making the business a potential botfarm or nest for spammers to attract unsuspecting customers. Unless the reviewer acknowledges the suspicious pricetag in their review, their score is deducted.

Average Rating: The product rating signals consumers how good the product is. A product can be good but not flawless; as such a product that has an average rating of anything more than 4.5/5 conveys to the annotators that several fake reviews have caused this bias to occur. So, the score is to be deducted.

Username: A username is a vital part of the review authenticity as it acts as a placeholder for the identity of the reviewer. People generally give a username that has derived meaning whereas bot farms that mass produce accounts tend to give it names that have no meaning like numbers, or a mishmash of letters that make no meaning. This indicates to the annotators that this may be a fake review from a bot account; As such the score is deducted.

Date: Bot farms or Spammers tend to mass-review a product in a single day to produce immediate results which is an unusual sighting that notifies the annotators that this may be a spammer[23]. Hence a deduction is made to the score.

Verified Purchase: A verified review is done when a reviewer purchases the product and the system matches the review of the product with a history of purchases from the said account. This is further proven by fake reviews that exemplify the product without any verification that the account that made the review has bought the product [24]. A verified review is much more reliable as we know the reviewer has bought and to some extent used the product. This is why the participants judge the verified review status as the norm and lack thereof directs this towards it being a dishonest review.

Number of Helpful Votes: Helpful votes are a metric in Amazon reviews that show how helpful a review was to other people. Most fake reviews focus on the features of the product whereas the average individuals detail their experience with the product as well as its flaws which are embellished with the human experience. That is to say that not having helpful votes is a negative. The annotators as such see the helpful votes as indicators of the authenticity of the reviews. Therefore, the score increases per the number of helpful votes.

Image Status: Images on reviews are a telltale sign of a product but, with the advent of bot farms they are made to increase online traffic with an agenda of falsehood. This is why the image is also an integral part of how much the average individual has invested at the bare minimum to prove its validity in the online world. So, we take having an image as the norm and not having an image with a negative connotation that the annotators implicitly understand.

Vine Review Status: Amazon Vine is an invite-only program designed to give customers products at a free or discounted price in exchange for honest and unbiased reviews. However, businesses may take advantage of the system by receiving biased reviews in exchange for their products; thus, conversely influencing the customer to give fake reviews, as such the annotators will deduct the score if they see a review done using Amazon Vine.

Review Details: The participant judges the review details on some criteria based on its length, the content of the product details, and its usage of words: whether it is incoherent, excitable, or repetitive even across the product reviews as online platforms often have a significant number of duplicate reviews with many being spam [2]. According to established papers, spam reviewers rarely write detailed reviews on the product to bypass spam detection [23]. However, the converse may also be true as of 2024 with the advent of AI technology. A lengthy review may be easily generated using tools readily accessible to the populace such as chatGPT. Moreover, fake reviews rarely discuss the product itself. Instead, these reviews may push consumers to purchase the product based on its features rather than forming an opinion on it [1], [24]. These reviews may also promote the product of competitors to harm the target business [37]. Subsequently, not only may they lack any relevant information they may also be hard to read or understand because of grammatical errors, unusual phrasing, or use of words [4]. Other than the general readability, reviews that emphasize excitement(or irritability) with the use of capitalized letters or exclamation marks, bring into question their authenticity. If any criterion matches, a deduction is made from the overall score.

Review-Rating Polarity: Occasionally, reviews may present a paradox where a poor rating is given despite the text praising the product [4]. The reverse can also be observed. Notably, fake reviews tend to exhibit extreme sentiment polarity, being either exceptionally positive or overwhelmingly negative [28]. In both cases, a deduction is made from the overall score.

3.2.2 Data Analysis

The dataset used in this study was sourced from the Amazon, comprising 12,915 reviews. Key features encompass product and review details, ratings, price, and verification status of the user. Employing a semi-supervised approach, we manually labeled 7,727 reviews with the assistance of university students. The dataset is partitioned into an 80:10:10 ratio, with 80% designated for training, 10% for testing, and another 10% for validation.

3.2.3 Peer review

Peer Review is the process of evaluating research papers using independent sources. To prevent bias, university students engaged in a peer review where each student evaluated a data section different from their initial training set, and vice versa. This ensured all sections received an independent and accurate assessment. The scores were then averaged to ensure the quality of the dataset.

3.2.4 Data Preprocessing

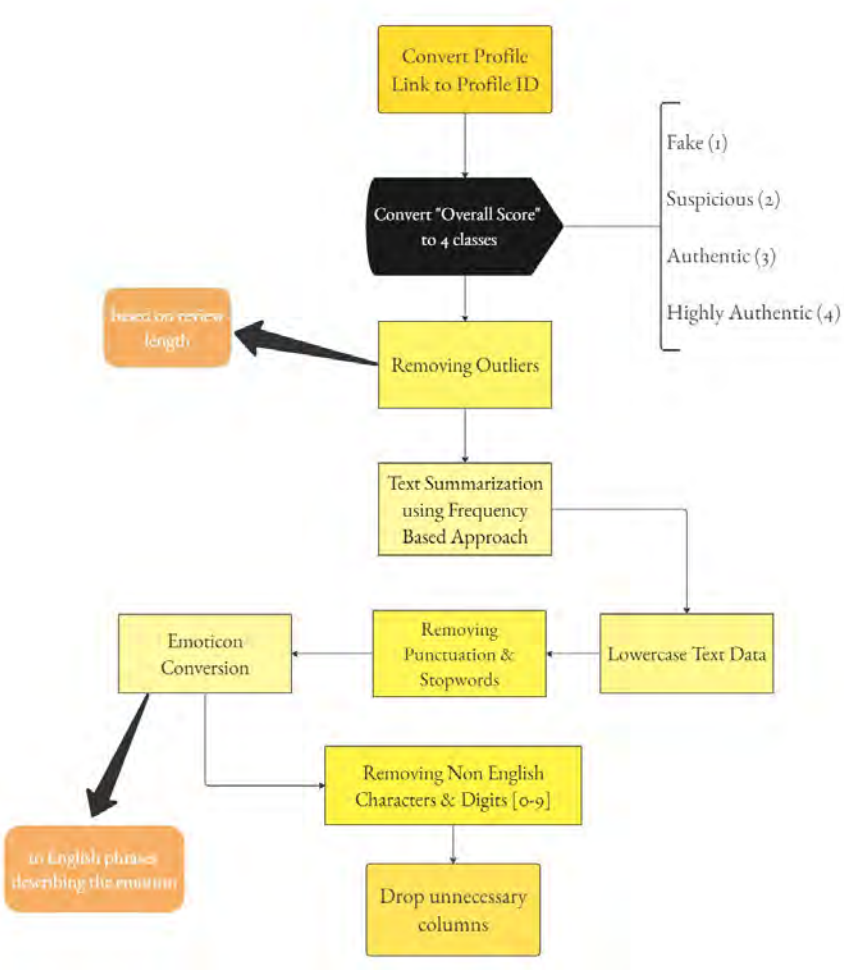


Figure 3.3: Preprocessing Workflow

Data preprocessing is the process of editing and removing all unnecessary values that do not impact the model and as such adapting some values that the model can comprehend and use adequately.

Initially, this process starts with the profile link converted into a profile ID due to some users remaining anonymous. This is done for anonymous Amazon and Kindle customers. Subsequently, we convert the overall score into 4 different categories that are fake, suspicious, authentic, and highly authentic.

Score	Label	Review Type
0 to 2.9	1	Fake
3 to 5.9	2	Suspicious
6 to 7.9	3	Authentic
8 to 10	4	Highly Authentic

Table 3.1: Score Classification Table

Proceeding with the classification, it can be observed that the dataset is imbalanced. The fake class (1) is the minority class.

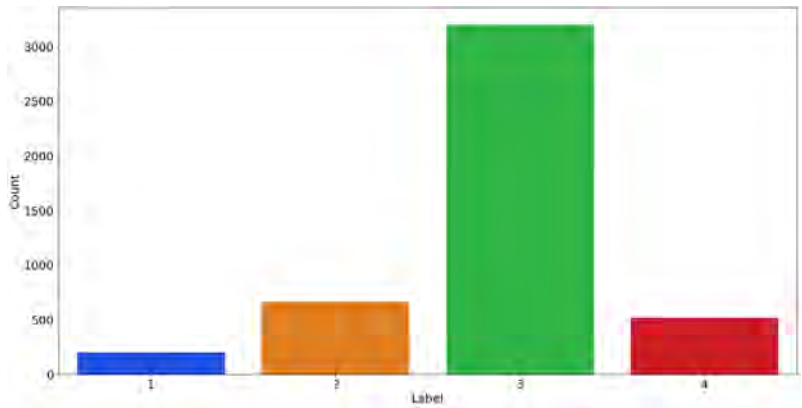
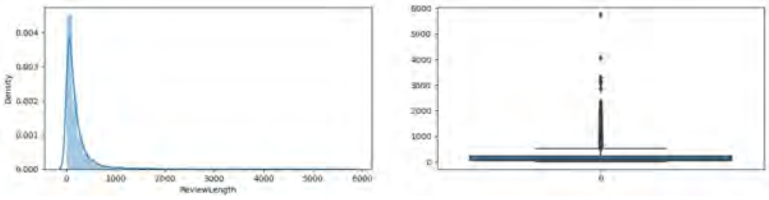


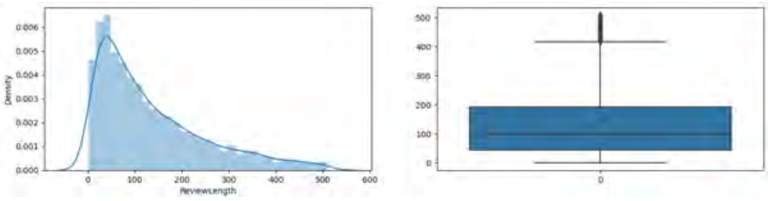
Figure 3.4: Data Distribution

Moreover, the text, punctuation, and stopwords are removed. Emoticon conversion is done to transform them into textual data to describe the emotions. For example: ":)" is converted into "I am happy." Non-English characters and digits [0-9] are removed from the dataset as well as. Then the product URL, review link, profile link, annotator comments as well as the now unnecessary overall score columns, empty and duplicate rows are dropped from the dataset.

All reviews exceeding the average length are removed using IQR-based filtering since outliers skew the model parameters and impact the overall result of the model. Fig. 3.5 demonstrates the distribution of review lengths before and after outliers have been removed.



Before Filtering Outliers



After Filtering Outliers

Figure 3.5: Outliers based on Review Length

Then we implement an abstract summarizer using a frequency-based approach to analyze the review details and create new phrases that convey the essence of the original text more concisely for all reviews. Otherwise, ConvBERT will truncate those that exceed 512 tokens and significant meaning will be lost. One hot encoding converts categorical data to binary values. This is done to avoid bias and to handle the categorical data without any meaning. Here, this is applied to the fake, suspicious, authentic, and highly authentic labels into their respective binary values. Finally, the ConvBERT encoder then tokenizes the textual data to produce input ID, attention mask, segment IDs, and numerical data.

3.3 Model

3.3.1 BERT

BERT which stands for Bidirectional Encoder Representations from Transformers, represents a novel approach to natural language processing (NLP) introduced by Google in 2018. Transformers uses self-attention to determine how words in a line or string are connected. BERT stems from this idea. Models before BERT read text unidirectionally; either from left-to-right or right-to-left. However, BERT is capable of reading from both sides simultaneously. Here are some of the most important facts on BERT:

1. **Pre-trained Representation:** BERT is trained on large sets of data, like Wikipedia pages and BookCorpus, using unsupervised learning. Preliminary training helps the model learn how to use words to describe different situations.

2. **Contextual Understanding:** BERT can comprehend by reading the whole sentence or sequence and putting words together in ways that make sense with that sentence or sequence. This makes the model smart about languages, like how some words can mean more than one thing and how their meaning changes depending on the situation.

Now, the language model uses two fundamental learning methods: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM masks input words and trains the algorithm to predict them based on the context the sentence provides. NSP on the other hand predicts if one sentence follows another in a document. From this challenge, BERT learns the relationships between sentences. After pre-training, BERT can be fine-tuned on smaller datasets designed for more specific tasks like text classification, named object recognition, and so on. When fine-tuning, the parameters are adjusted such that they fit the task.

Another key component of BERT is self-attention. Each word in a sentence starts as an embedding vector. These embeddings are then transformed into three vectors: Query Q, Key K, and Value V. Taking the dot product of the Query of the first word and the Key of the second word, the attention score can be calculated. In this manner, the relationship between that word and every other word in the sentence can be captured. Following this, the attention weights are found through scaling and the application of a softmax function. The attention weights are then used to compute a weighted sum of the Value vectors V of all the words in the sentence.

Mathematically, it can be expressed as

$$\text{Self-Attn }(Q, K, V) = \text{SoftMax}((QKT) \;/\; \sqrt{dk})\text{V}$$

Here Self-Attn denotes self-attention.

The key ability of BERT lies in its self-attention mechanism, often called global self-attention. This part of the model allows it to pay attention to any part of the input sequence, no matter how far apart they are. The traditional self-attention system lacks an understanding of word order. In compensation, BERT utilizes positional encodings. Positional encodings, added to word embeddings, inform BERT about the word order, enabling its bidirectional self-attention mechanism to attend to each word from both sides simultaneously. This facilitates the model with a deeper contextual understanding.

3.3.2 ConvBERT

Convolutional BERT or **ConvBERT** for short, is a small and efficient iteration of BERT introduced to counter the increased memory footprint and computational costs of the original model due to global self-attention.

Since global self-attention lets each word in the input sequence attend to every other word, irrespective of their distance it consequently has some drawbacks including redundant attention that contributes to overhead without substantial benefit; as words span greater distances, their impact on the final output diminishes. Similarly, computing attention weights become more complex with longer sequences. As a result, the model may incur a larger memory footprint and computational cost due to its need to store more intermediate values. Study [14] demonstrated this redundancy by disabling some attention heads in both pre-trained and fine-tuned BERT which surprisingly, did not always result in a decline in accuracy but rather an increase in performance.

To address these issues, [22] proposed a novel technique: span-based dynamic convolution. This approach replaces some attention heads with convolution heads, as convolution layers excel at extracting local features. The combination of convolution heads and remaining self-attention heads creates a mixed attention block capable of both local and global learning.

To implement local self-attention, dynamic convolution replaces standard convolution. This allows the generation of more adaptive convolution kernels by considering input spans rather than individual tokens. This improves the flexibility of the model as it generates kernels for the same input in different contexts. Fig. 3.6 demonstrates how the same token with its direct neighbors in other words the the local span, generates different kernels based on context. The first 'saw' is a verb whereas the second 'saw' is a noun hence why the produced kernels are separate in appearance.

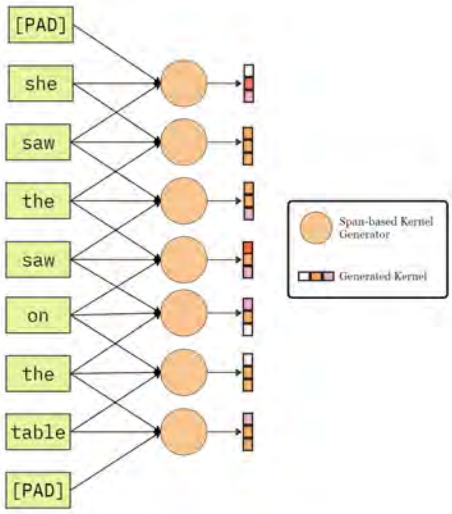


Figure 3.6: Span-based Dynamic Convolution

Furthermore, tokens are transformed into a smaller hidden representation by projecting input tokens to a lower-dimensional space. The feed-forward module in BERT consists of two fully connected linear layers with an activation set in between and is adjusted to a higher dimensionality in the inner layer to capture complex word relationships. However, this adjustment increases model size, training time, and in turn the computational costs. In response, the grouped linear operator is introduced, dividing input and output vectors into groups. Instead of a single large matrix multiplication, smaller group-wise multiplications are performed, reducing parameters and connections in each group. This innovation gives rise to the compact model ConvBERT.

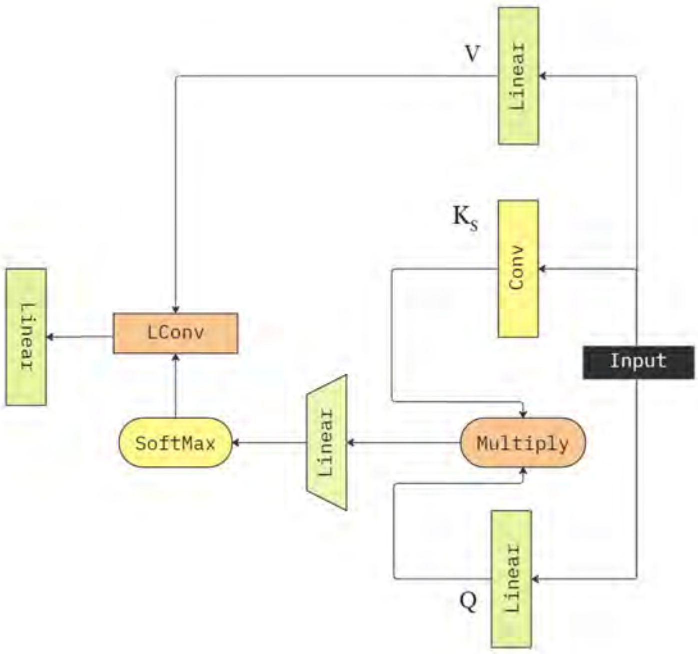


Figure 3.7: Span-based Dynamic Convolution Structure

We represent span-based dynamic convolution SDConv as

$$\text{SDConv}(Q, K_s, V; W_f, i) = \text{LConv}(\text{V}, \text{Softmax}(W_f(Q \odot K_s)), i)$$

Here a position-dependent kernel, $W = f(Xi)$, is dynamically created for each position i using a linear model f with learnable weight W_f. This kernel, along with the Query Q, Span-aware Key K_s, and Value V, is employed within a lightweight convolution $LConv$. Fig. 3.7 illustrates the structure of span-based dynamic convolution.

The mixed attention block combines both self-attention and the span-based dynamic convolution to capture both global and local dependencies with less redundancy. We represent mixed attention Mixed-Attn as

$$\text{Mixed-Attn}(K, Q, K_s, V; W_f) = \text{Cat}(\text{Self-Attn}(Q, K, V), \text{SDConv}(Q, K_s, V; W_f))$$

Here Cat denotes Concatenation.

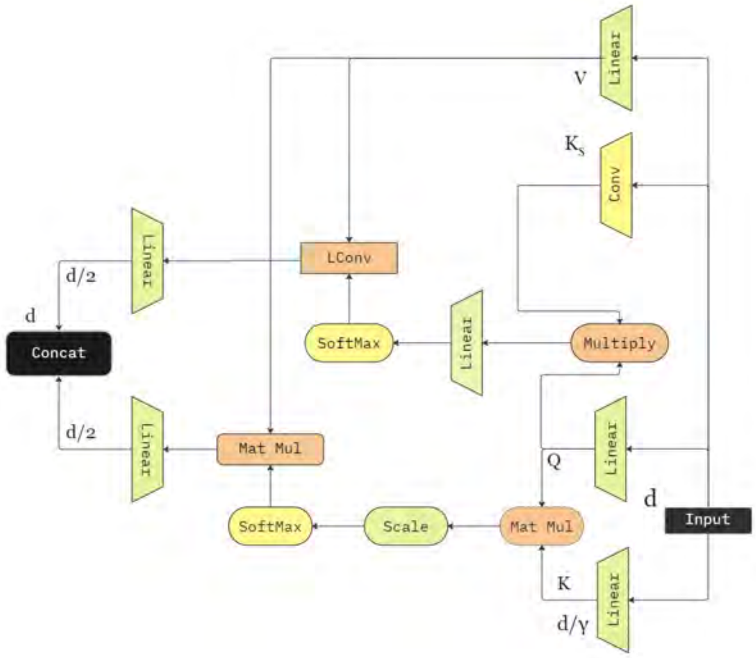


Figure 3.8: Mixed Attention Block Structure

BERT typically uses hefty d-dimensional word embeddings and multiple attention heads. ConvBERT reduces the embedding space to $d\backslash\gamma$ and reduces the number of attention heads by γ ratio where $\gamma > 1$. This optimizes the transformer for efficiency while still preserving its effectiveness. Fig. 3.8 illustrates the mixed attention block.

3.3.3 LSTM

The **LSTM** or Long Short-Term Memory model is a type of recurrent neural network (RNN) meant to fix the disappearing gradient problem common in RNNs. It is effective in capturing and processing sequential data, making it well-suited for various NLP tasks. To begin with, Sepp Hochreiter and Jürgen Schmidhuber invented LSTMs in 1997. The main difference between RNNs and LSTMs is that LSTMs have a gated unit, or cell, as their hidden layer. In each cell, four layers work together to create its output and state.

LSTM networks consist of three gates: the Forget Gate, which stores or discards data; The Forget Gate takes two inputs, x_t (input at the current time) and h_{t-1} (output from the preceding cell), multiplies them using weight matrices, and then adds bias. An activation function is applied to the output to produce a binary value. When the output for a specific cell state is 0, the data is erased from memory, but when the output is 1, the data is stored for later use. The input gate, adds new data. First, with inputs h_{t-1} and x_t, the data is sigmoid-filtered to determine which values to remember, much like a Forget Gate. Next, the tanh function is used to obtain an output ranging between -1 to +1. This vector will contain all the possible values for h_{t-1} and x_t. We then multiply the vector values by the regulated values and the Output Gate, which uses the current cell state to establish the next concealed state. The tanh function is applied to the cell to build a vector first. Inputs h_{t-1} and x_t are used to filter the data, and the sigmoid function is used to regulate it based on the values to be remembered. Finally, to send the contents of the vector and the controlled values to the next cell, they are multiplied.

3.3.4 BiLSTM

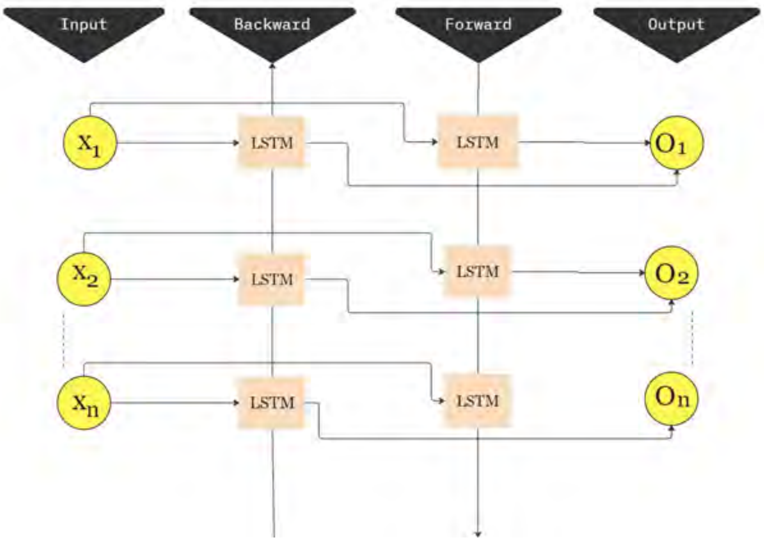


Figure 3.9: BiLSTM Structure

Bidirectional LSTM (**BiLSTM**) is a sequence model with two LSTM layers: one processes input in the forward direction, and the other in the backward direction, handling the input sequence in both ways. This bidirectional processing, enabled by two sets of different hidden states and cell states, captures information from past and future contexts. It outperforms traditional unidirectional LSTMs by simultaneously processing sequences in both directions, improving the understanding of temporal dependencies, and handling of long-range dependencies. Particularly effective in NLP tasks, BiLSTM is adept at mitigating the vanishing gradient problem accurately, contributing to stable training in scenarios with long-term dependencies and making it suitable for various tasks requiring accurate predictions. Its versatility makes it suitable for various tasks requiring accurate predictions. For a deeper understanding of the BiLSTM, let us examine the computation process of the LSTM unit in the following steps:

1. W_c, b_c, and $\widetilde{C}_t$ denote the weight matrix, bias, and candidate memory respectively.

$$\widetilde{c}_t = \tanh_{(W_c.[h_{t-1},x_t]+b_c)}$$

2. i_t, b_i, and W_i denote the input gate, its bias, and its weight matrix respectively.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$$

3. f_t and W_f denote the forget gate and its weight matrix respectively.

$$f_t = \sigma(W_i.[h_{t-1}, x_t] + b_f)$$

4. c_t denotes the memory cell current moment.

$$c_t = f \times c_{t-1} + i_t * \widetilde{c}_t$$

5. o_t denotes the output gate.

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0)$$

6. h_t denotes the output of an LSTM unit.

$$h_t = o_t * \tanh c_t$$

3.3.5 Proposed Model

Our proposed model joins the ConvBERT with two stacked BiLSTMs. As mentioned earlier, the ConvBERT tokenizer returns four outputs, Input ID, Attention Mask, Segment ID, and Numerical data. The first three textual outputs, Input ID, Attention Mask, and Segment ID are computed in the ConvBERT layer. Separated, the numerical data has a substantial weight for classification. It is passed through a dense layer followed by a dropout layer to prevent overfitting. The product of the ConvBERT layer and dropout layer are concatenated which creates our initial startup for the BiLSTM layer. The first BiLSTM layer returns the sequences. The second layer processes those sequences and only returns a single vector output containing information on the initial input sequences. The vector output is then further processed through the dense-dropout-dense layers.

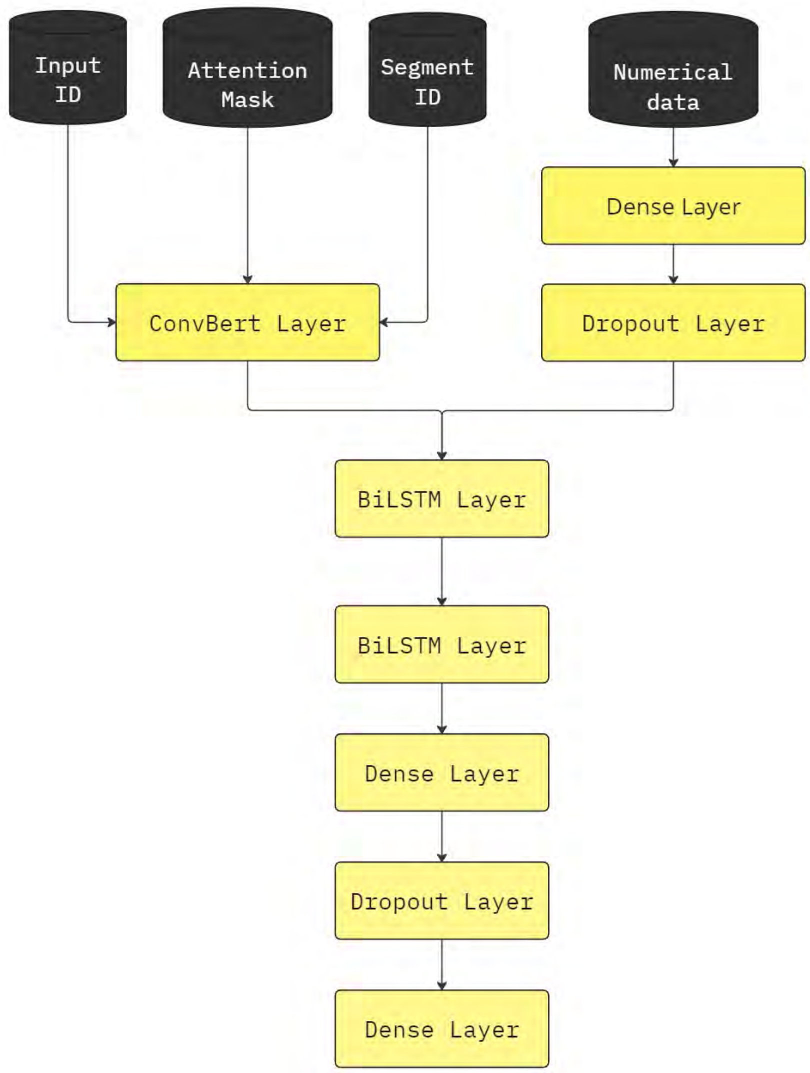


Figure 3.10: ConvBERT-BiLSTM Structure

Chapter 4

Results & Analysis

4.1 Experimental Setup

In our experimental configuration, we harnessed a computational system featuring an Intel Core i5-10400F processor, 16GB of RAM, and a high-performance GPU, either the NVIDIA GeForce RTX 3070 or 3080 Ti. For optimization, we utilized the Adam optimizer with a learning rate set to $1e^{-5}$. During training, a batch size of 8 was applied, and the models underwent 35 epochs. The neural architectures included ConvBERT, ALBERT, ELECTRA, and Flan T5. Additionally, we implemented dynamic learning rate adjustment with the ReduceLROnPlateau technique and early stopping for efficient model training. The evaluation metrics calculate the F1 score accuracy and weighted average.

4.2 Experimental Results

Model	Weighted Average			Accuracy
	Precision	Recall	F1 Score	
ConvBERT-BiLSTM	0.97	0.97	0.97	0.97
ALBERT	0.92	0.89	0.90	0.89
ELECTRA	0.92	0.90	0.90	0.90
Flan T5	0.81	0.67	0.70	0.67

Table 4.1: Performance Report

Table 4.1 summarizes how well the proposed model and reference models: ALBERT, ELECTRA, and Flan-T5 perform on the classification task, including their precision, recall, F1 score, and accuracy.

Precision: Precision gauges the accuracy of the positive predictions of the model. It is the ratio of true positives to all predicted positives.

$$Precision = \frac{TP}{(TP + FP)}$$

Recall: Recall measures the ability of the model to correctly identify actual positives. It is the ratio of true positives to all actual positives in other words: Both true positives and false negatives.

$$Recall = \frac{TP}{(TP + FN)}$$

F1 Score: The F1 score balances precision and recall using their harmonic mean.

$$F1 = \frac{2(Precision * Recall)}{(Precision + Recall)}$$

Accuracy: Accuracy assesses the overall correctness by calculating the ratio of correct predictions to the total predictions.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Weighted Average: For a multi-class system, the weighted average computes the metric means for each class, with weights based on class support (number of instances). This accounts for class imbalances. In other words, for a critically imbalanced dataset such as ours, the weighted average is a suitable measure for evaluation.

Model	Parameters
ConvBERT-BiLSTM	109,521,676
ALBERT	11,851,652
ELECTRA	109,059,716
Flan-T5	109,796,612

Table 4.2: Model Parameters

Table 4.2 provides the parameters used for each model. ConvBERT-BiLSTM, ELECTRA, and Flan-T5 have similar parameters, approximately 109M each. However, ALBERT utilizes significantly fewer parameters, about 9 times less than the other models.

4.3 Performance Analysis

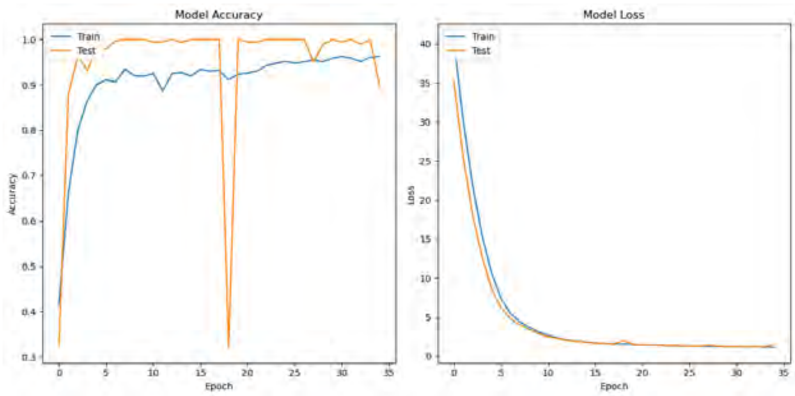


Figure 4.1: Train-Validation & Accuracy-Loss: convBERT-BiLSTM

From Fig. 4.1 a rapid rise can be observed. We presume it to be due to the number of layers, allowing the model to pick up on patterns rather quickly. Following the rise, hovering over 90 percent, the curve becomes constant. Notice the sharp fall between the 15th and 20th epochs: There is a shift in weights from the previous epoch to the next which causes the model to make a faulty prediction despite it being very close to the correct answer. Hence the validation accuracy drops. However, the loss curve is hardly affected as there is no overfitting.

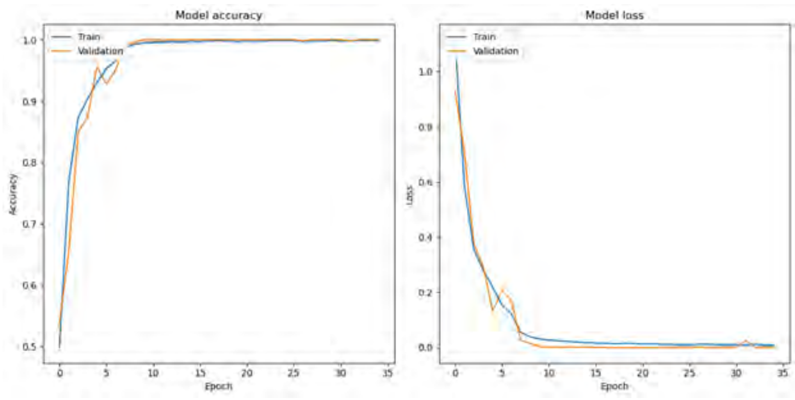


Figure 4.2: Train-Validation & Accuracy-Loss: ALBERT

There is little fluctuation in Fig. 4.2 in other words ALBERT is good with small data samples to train on since its parameters are lower than the other models. The low fluctuation rate is also an indication that the learning rate is suitable for ALBERT.

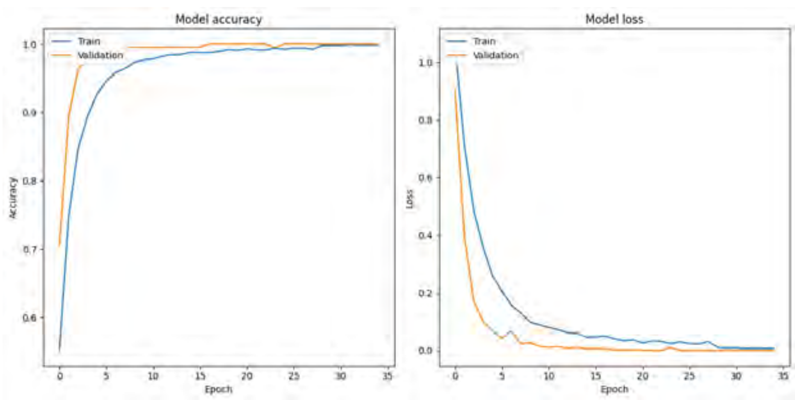


Figure 4.3: Train-Validation & Accuracy-Loss: ELECTRA

ELECTRA performs similarly to ALBERT as shown in Fig. 4.3. However, because it has more parameters it requires more time to process the sequences.

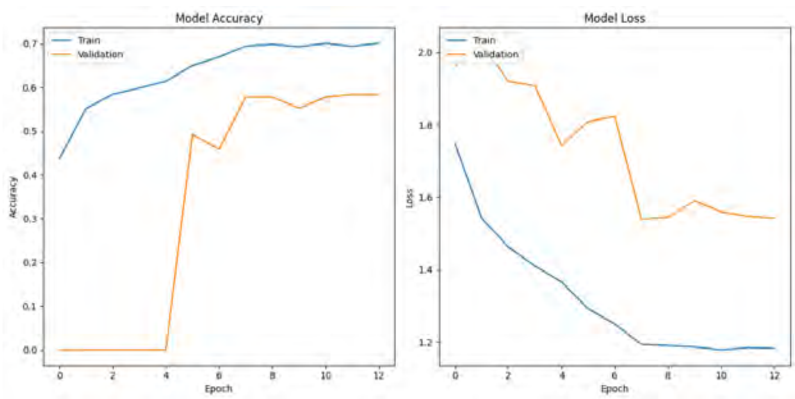


Figure 4.4: Train-Validation & Accuracy-Loss: T5

The Flan T5 is the oddball reference model. As you can see in Fig. 4.4 the model only ran 12 epochs, issued by early stopping. Overfitting occurs because, unlike the BERT models, T5 models are mainly used for text generation. Although it is capable of text classification tasks, the train settings prevent it from doing well. To maintain a constant environment to base performance comparison based on model architecture, the layer density, layer count, and other hyperparameters remained unchanged across all models.

Table 4.1 reports the accuracy and weighted average for each metric for all the models. The proposed model, ConvBERT-BiLSTM has the highest accuracy and weighted F1 average, both amounting to 0.97. Both ALBERT and ELECTRA demonstrate the same weighted F1 averages (0.90) and are almost tied in their accuracy (0.89 and 0.90 respectively). Then again in terms of accuracy-to-parameter ratio, ALBERT is most efficient.

Chapter 5

Conclusion

In conclusion, the credibility of customer reviews in the online marketplace is undetermined by unethical practices such as review manipulation. Although various detection models have been studied, there remains a pressing need for comprehensive surveys in this field. To address this gap, we have hybridized ConvBERT with BiLSTM which has outperformed ALBERT, ELECTRA, and Flan-T5, provided that the batch size, epochs, and learning rate were kept constant across models for fair comparison in terms of model architecture rather than train settings. Our proposed model attained an accuracy of 97%, surpassing the other models by a landslide. However, ALBERT is most efficient when considering the accuracy-to-parameter ratio. These results and analysis may prove beneficial to businesses and consumers alike.

In the future, further improvements to the rubric may be necessary for the assurance of quality data. We plan to develop a full-fledged software that runs as a browser extension on popular e-commerce websites. The extension will have a profile-based feature where the neural network is capable of telling genuine users apart from their profiles. Last but not least, we believe these insights will prove valuable for buyers when making informed decisions on which products to purchase and which of them to avoid.

Bibliography

[1] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, IEEE, 2007, pp. 547–552.

[2] ——, "Review spam detection," in *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada: ACM, Oct. 2007, pp. 1189–1190.

[3] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam.," Jan. 2011, pp. 2488–2493. DOI: 10.5591/978-1-57735-516-8/IJCAI11-414.

[4] T. Lappas, "Fake reviews: The malicious perspective," in *Natural Language Processing and Information Systems*, G. Bouma, A. Ittoo, E. Métais, and H. Wortmann, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 23–34, ISBN: 978-3-642-31178-9.

[5] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," UIC-CS-03-2013, Technical Report, 2013.

[6] Y. Li, X. Feng, and S. Zhang, "Detecting fake reviews utilizing semantic and emotion model," in *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*, 2016, pp. 317–320. DOI: 10.1109/ICISCE.2016.77.

[7] L. Chen, T. Jiang, W. Li, S. Geng, and S. Hussain, "Who should pay for online reviews? design of an online user feedback mechanism," *Electronic Commerce Research and Applications*, vol. 23, pp. 38–44, 2017, ISSN: 1567-4223. DOI: https://doi.org/10.1016/j.elerap.2017.04.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1567422317300224.

[8] J. Fontanarava, G. Pasi, and M. Viviani, "Feature analysis for fake review detection through supervised classification," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, pp. 658–666. DOI: 10.1109/DSAA.2017.51.

[9] N. Nguyen. (May 2018). "Here's another kind of review fraud happening on amazon," [Online]. Available: https://www.buzzfeednews.com/article/nicolenguyen/amazon-review-reuse-fraud.

[10] K. Taşağal and Ö. Uçar, "Detection of fake user reviews with deep learning," *International Journal of Research in Engineering and Applied Sciences (IJREAS)*, vol. 8, no. 12, Dec. 2018. [Online]. Available: https://ssrn.com/abstract=3319640.

[11] Y. Wang, X. Lu, and Y. Tan, "Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines," *Electronic Commerce Research and Applications*, vol. 29, pp. 1–11, 2018, ISSN: 1567-4223. DOI: https://doi.org/10.1016/j.elerap.2018.03.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1567422318300279.

[12] A. Heinzman. (Mar. 2019). "How fake reviews are manipulating you online," [Online]. Available: https://www.howtogeek.com/407521/how-fake-reviews-are-manipulating-you-online/.

[13] R. P. Kashti and P. S. Prasad, "Analysis of classifiers for fake review detection," *International Journal for Technological Research in Engineering*, vol. 6, no. 9, pp. 5414–5418, May 2019, ISSN: 2347-4718. [Online]. Available: https://ijtre.com/wp-content/uploads/2021/10/2019060902.pdf.

[14] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, *Revealing the dark secrets of bert*, 2019. arXiv: 1908.08593 `[cs.CL]`.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, 2019. arXiv: 1907.11692 `[cs.CL]`.

[16] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, *Electra: Pre-training text encoders as discriminators rather than generators*, 2020. arXiv: 2003.10555 `[cs.CL]`.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, *Albert: A lite bert for self-supervised learning of language representations*, 2020. arXiv: 1909.11942 `[cs.CL]`.

[18] D. Proserpio, B. Hollenbeck, and S. He. (Nov. 2020). "How fake customer reviews do — and don't — work," [Online]. Available: https://hbr.org/2020/11/how-fake-customer-reviews-do-and-dont-work.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*, 2020. arXiv: 1910.01108 `[cs.CL]`.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *Xlnet: Generalized autoregressive pretraining for language understanding*, 2020. arXiv: 1906.08237 `[cs.CL]`.

[21] T. Fornaciari, F. Bianchi, M. Poesio, and D. Hovy, "BERTective: Language models and contextual information for deception detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 2699–2708. DOI: 10.18653/v1/2021.eacl-main.232. [Online]. Available: https://aclanthology.org/2021.eacl-main.232.

[22] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, *Convbert: Improving bert with span-based dynamic convolution*, 2021. arXiv: 2008.02496 `[cs.CL]`.

[23] R. Mohawesh, S. Xu, S. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, Apr. 2021. DOI: 10.1109/ACCESS.2021.3075573.

[24] B. Stegner, "8 ways to spot fake and useless reviews online," *[Website not provided]*, Oct. 29, 2021. [Online]. Available: https://www.makeuseof.com/tag/how-to-spot-fake-reviews/.

[25] Amazon Staff. (Jul. 2022). "Amazon targets fake review fraudsters on social media," [Online]. Available: https://www.aboutamazon.com/news/policy-news-views/amazon-targets-fake-review-fraudsters-on-social-media.

[26] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pre-training and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. DOI: 10.18653/v1/2022.findings-naacl.98. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98.

[27] M. Brewster. (Apr. 2022). "Annual retail trade survey shows impact of online shopping on retail sales during covid-19 pandemic," [Online]. Available: https://www.census.gov/library/stories/2022/04/ecommerce-sales-surged-during-pandemic.html.

[28] M. Clarke. (Sep. 2022). "How to create a fake review detection model," [Online]. Available: https://practicaldatascience.co.uk/machine-learning/how-to-build-a-fake-review-detection-model.

[29] P. Gupta, S. Gandhi, and B. R. Chakravarthi, "Leveraging transfer learning techniques- bert, roberta, albert and distilbert for fake review detection," in *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, ser. FIRE '21, Virtual Event, India: Association for Computing Machinery, 2022, pp. 75–82, ISBN: 9781450395960. DOI: 10.1145/3503162.3503169. [Online]. Available: https://doi.org/10.1145/3503162.3503169.

[30] R. Kanmani and S. B, "Leveraging readability and sentiment in spam review filtering using transformer models," *Computer Systems Science and Engineering*, vol. 45, Nov. 2022. DOI: 10.32604/csse.2023.029953.

[31] J. Salminen, C. Kandpal, A. M. Kamel, S.-g. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102 771, 2022, ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2021.102771. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969698921003374.

[32] S. Saumya and J. P. Singh, "Spam review detection using LSTM autoencoder: An unsupervised approach," *Electronic Commerce Research*, vol. 22, no. 1, pp. 113–133, 2022. DOI: 10.1007/s10660-020-09413-4.

[33] A. Schuman. (Nov. 2022). "What fake reviews can mean for your business," [Online]. Available: https://www.bazaarvoice.com/blog/what-fake-reviews-can-mean-for-your-business/.

[34] S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun, "Scoregan: A fraud review detector based on regulated gan with data augmentation," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 280–291, 2022. DOI: 10.1109/TIFS.2021.3139771.

[35] M. Stieb. (Jul. 2022). "Amazon's war on fake reviews," [Online]. Available: https://nymag.com/intelligencer/2022/07/amazon-fake-reviews-can-they-be-stopped.html.

[36] N. Wang, J. Yang, X. Kong, and Y. Gao, “A fake review identification framework considering the suspicion degree of reviews with time burst characteristics,” *Expert Systems with Applications*, vol. 190, p. 116 207, 2022, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.116207. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421015219.

[37] S. Hillgear, “How to spot fake reviews on amazon,” Jul. 8, 2023. [Online]. Available: https://www.wired.com/story/how-to-spot-fake-reviews-amazon/.

[38] M. Loiselle. (2023). “3 statistics that show how customer reviews influence consumers,” [Online]. Available: https://www.dixa.com/blog/3-important-statistics-that-show-how-reviews-influence-consumers/.

[39] A. Q. Mir, F. Y. Khan, and M. A. Chishti, *Online fake review detection using supervised machine learning and bert model*, 2023. arXiv: 2301.03225 `[cs.CL]`.

[40] G. M. Shahariar, M. T. R. Shawon, F. M. Shah, M. S. Alam, and M. S. Mahbub, *Bengali fake reviews: A benchmark dataset and detection system*, 2023. arXiv: 2308.01987 `[cs.CL]`.

[41] B. Yılmaz. (Jan. 2023). “Fake review detection in 2023: Overview, methods & case studies,” [Online]. Available: https://research.aimultiple.com/fake-review-detection/.