National College of Ireland

Identifying Fake Product Reviews

MSc Research Project
MSc Data Analytics

Mounika Yarlagadda
Student ID: X20140924

School of Computing
National College of Ireland

Supervisor: Mohammad Hasanuzzaman

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Mounika Yarlagadda

Student ID: X20140924

Programme: MSc Data Analytics **Year:** 2022

Module: Research Project

Supervisor: Mohammad Hasanuzzaman

Submission Due Date: 31st January

Project Title: Identifying Fake Product Review

Word Count: 7175

Page Count: 22

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: Mounika Yarlagadda

Date: 31/01/2022

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):	

Identifying Fake Product Reviews

Mounika Yarlagadda
X20140924

Abstract

Fake review detection can help us determine if a product review is real or false. Different methods can be used to achieve this, but we will focus on getting the result using machine learning methods. Our approach focuses on the content of the user review text. People who write fake reviews often choose topics or words to impress online customers, so their choice of words will be different from others. This word selection can be used to distinguish between false and false reviews. If these updates are received correctly, fake updates can be automatically removed once detected, which helps to provide only factual information and more specifically to companies and markets to customers.

The aim of the research is to develop an online technology system to detect and eliminate fake reviews with the aim of protecting the interests of customers, products, and e-commerce portals. The Flipkart Review dataset is analyzed with the help of Natural Language Processing, Supervised Learning Model and Deep Learning Model. Data was collected from a single shopping website (Flipkart) to identify counterfeit product reviews.

1 Introduction

1.1. Background Scope

As the world changes with the advent of technology the opportunities for consumers to buy goods online are also increasing. The growing trend of online shopping has forced many people to read reviews before returning to buy a product online and return their opinion. Companies are increasingly affected as online shopping has a huge impact on the growing economy. Therefore, the chances of spam views are increasing these days as many customers rely on online updates. The effect of such information on the organization and its customers is to encourage people to improve or reduce productivity. Positive reviews have had a significant impact on the reputation and reputation of companies. However, negative perceptions can cause significant social and economic harm. A company hires people to write positive reviews or negative reviews from competitors about their products.

The invention of these illegal updates has become an important issue to provide relevant and useful information. Native language learning is concerned with how computers can help humans understand language, text, and speech. It can be used to understand things practically. False updates can be detected through natural language analysis using a variety of machine learning methods. This review helps to capture details and patterns from the content of the text. Data helps to compare various updates and detect fraud. Users who write false reviews choose different words or patterns to impress others. This can be used as a way to get illegal and false updates. So far three types of false reviews have been identified,

1) False comments
2) Various product reviews
3) Non-updated containing various advertisements and links.

False reviews are very embarrassing because they undermine the integrity of the online review system. It's hard to tell the difference between false and true reviews when it comes to personal reading, so it's very difficult to find these spam reviews. Consumer reviews play an important role in understanding the specific market situation of products and companies.

Decision systems solve rapidly changing and sophisticated problems and find solutions. Existing electronic teaching methods can be divided into supervised and indirect methods. Second, they can be divided into three categories based on their characteristics: character, language, or a combination of the two.

1.2. Motivation

As the consumption of online products increases rapidly, the competition in the market is increasing day by day. Business executives can market their products to outsiders by promoting and discrediting other competing products or hiring them to post and deliver illegal judgments about products. Therefore, different methods and techniques are required to protect the authenticity of online products and ideas.

The financial evaluation of the status, views and reviews of various products can be very costly for any one company based on those reviews. If the review is true and true, the results will appear and be correct, but the situation will come back if we meditate on the other side. False reviews can lead to huge corporate losses, resulting in better and better products being sold, costing the developing company a lot of money.

1.3. Research Questions

Research questions describe the essence of the research project. Research questions can help in conducting detailed studies for counterfeit product review analysis. Machine learning techniques are used for predictive analysis and research project studies.

1) What is the rationale behind choosing a behavioral approach rather than a textual analysis approach?
2) What is the feature selection technique for the suggested approach?

1.4. Research Objective

Human decision making is one of the most complicated aspects in business operations. The current research emphasizes on the use of machine learning algorithms to predict the human decision-making approach and purchase decisions. The research focuses on the behavioral concept, over the textual approach, which allows the business organizations to detect the behaviors of the consumers, through the identification of the contents of the posts, views and reviews shared.

The technology helps in studying the way, how human beings exhibit a particular behavior, concerning the purchase of a specific product or service. The use of machine behavior allows the technical expert to research and evaluate on how the machines acquire or develop a specific individual or collective behavior (Singh and Tucker, 2017). It can be manifested in terms of developing a comprehensive focus, which centers on the concept of identifying the results, pertaining to the adaptability of the current business organization (Choudhury and Nur, 2019). In the present business environment, it is evident that the use of a hybrid approach is generally used to anticipate the human decision-making process.

In the given research, the focus is on identifying the falseness and truth of the reviews and posts, shared by the consumers, which ultimately lead to their decision-making process, along with influencing the users and others, who view the reviews (Mullainathan and Spiess, 2017). Hence, the use of the hybrid approach to Behaviour-Based Machine-Learning framework helps in recognizing the psychological features are essential for the representation of the data, reflecting upon the psychological properties underlying the competition baseline framework.

1.5. Structure of report

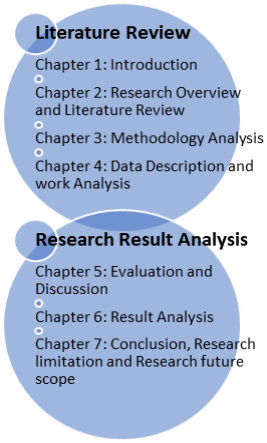

Figure 1: Report Structure

(Source: Self-created)

2. Related work

2.1. Introduction

This chapter of the research paper evaluates the different sources of information that is derived from the existing database, in regards to the selected subject of using Machine Learning for the reviewing fake detection system. Considering the growing significance of big data in today's business environment, it has become important to develop infrastructures and frameworks, which are integrated in form and ensures a closer and secured detection of data set privacy and security. The chapter identifies the fundamental points, which revolve round the subject, establishing the conceptual framework, by developing a relationship between the selected variables.

2.2. Significance of Machine Learning in current business operations

The contemporary business environment is characterized by exponential strategies to depend on technology, followed by adjusting with the external and internal factors that set the trend for the business operations. The

changes in the business operations tend to represent the growing significance of data science technology and use of big data to carry on with the business operations (Cavalcante *et al*.2019). Therefore, dealing with big data and addressing the issues of data breach is an integral part of current business operations.

Considering these changes in the current business environment, the use of machine learning is fundamental, at present. The use of machine learning facilitates effective data security and management and ensures that the processes of dealing with data challenges, can be conveniently dealt with the use of the machine learning technology (Leo *et al*.2019). It is evident from the present business environment that the growing importance of data science, has led to the identification of emphasizing and dealing with functions that help in harnessing the business operations and catalyzes the use of Big Data, Machine Learning and Artificial Intelligence, with each having its own respective focus and expertise of operation (Brunton *et al*.2020).

The use of machine learning out of all these technologies help in acknowledging this technology to enable business organizations to effectively gain insights from raw data. With the recognition of the machine learning algorithms, it needs to be stated that these algorithms can be used to iteratively learn from a given data set, understand patterns, behaviors, etc., all with little to no programming (Dogru and Keskin, 2020). These iterative processes, contribute towards development of a process, which is constantly evolving and developing, in respect to allowing the business organizations to ensure that they are always up to date with business and consumer needs (Kraus *et al.*2020). plus, it's easier than ever to build or integrate ML into existing business processes since all the major cloud providers offer ML platforms.

Hence, it is obvious that the use of machine learning cannot be denied from the current business environment, at any given point of time. The use of this method helps in developing a greater adaptability, with greater business resilience and enable more consumption of the accurate and more appropriate form of using analytics and predictions in regard to the views and opinions of the concerned client (Dou *et al*.2020). The use of machine learning algorithms, help in developing a more resilient and adaptable business framework, to the concerned business environment, followed by improvising the business operations and gaining a greater understanding about the current consumer preferences and the business operations (Kaissis *et al*.2020).

The use of machine learning is mostly in combination with the artificial intelligence and facilitates the development of a competent analytical process, leading to the identification of a more resilient framework. The machine learning services, like Amazon Sage Maker and Azure Machine Learning, has enabled users to utilize the power of cloud computing and integrate the same process, with the algorithms of machine learning, leading to the recognition of the needs of the business system (Dargan *et al*.2020).

The uses of machine learning algorithms in the current business operations is significant in terms of effective administering and managing big data, enabling the business organizations to understand the leading influencing factors and ensure that the processes are competent to deal with the difficult business environment (Yavanoglu and Aydos, 2017). Some of the key benefits of using the machine learning algorithm are shared below:

2.2.1. Facilitating User Behaviour analysis

The use of machine learning is extensively used in understanding the behaviour of the users and target audiences of the application. For instance, while running businesses, it is important to understand the attributes of the stakeholders, who are directly and indirectly associated with the business operations (Suthaharan, 2016). Hence, while running businesses, the companies tend to collect larger amounts of consumer information and running and managing this data, through machine learning algorithms, allows the business organizations to predict consumer purchasing habits, market trends, popular products, helping the business organizations to take the decisions, which are accurate and appropriate for the running of the business, in the most effective manner (Shang and You, 2019).

The use of this technology helps in streamlining the ordering according to market and consumer demand and even understands the logistical and operational processes (Boutaba *et al.*2018). It also helps in analysing the browser habits, through the automatic inclusion of the user data, followed by predicting the user experience and offers the targeted suggestions.

2.2.2. Fosters improved business automation

The business automation can be enhanced through the use of machine learning algorithms. The use of machine learning algorithms helps in evolving the business resources and operations and can be used in understanding the data, which is associated with the business operations (Yang *et al*.2015). This mechanism of machine learning framework enables in interpreting the current manufacturing models and recognizes all the deficiencies and gaps in the process (Raschka, 2015). Thus, it can be stated that the use of machine learning operates, beyond the conventional industrial operations and can improve the operational efficiency and predicting and deciphering the different data sets, associated with the operations.

2.2.3. Helps in recognizing cognitive responses and services

Machine learning also helps in recognizing and improving cognitive responses and services. This process is powered by the mechanisms of image recognition, followed by enabling the business organizations to develop a more secured environment and convenient authentication. With the help of the Natural Language Processing (NLP), the use of machine learning algorithm facilitates recognition of the different sets of clients and target audiences (Hasan *et al.*2019).

2.3. Importance of recognizing fake detection system

It is obvious that the current business environment is undergoing multiple changes, which are considerably affected and influenced by external and internal factors. It is one of the growing trends in the field of artificial intelligence, wherein it facilitates the recognition of deeper learning scenarios and allows the machines to acquire and read the data from the previous cases scenarios (Wang *et al*.2019). This allows the software professionals and technical experts to recognize the gaps in the process of technical learning and development, by using the data for future prediction and evaluation.

This identified variable in the current literature review, acknowledges the identification and review of fake detection systems, along with removing any possibility of fake reviews, especially in online reviews (Dey *et al*.2018). The introduction chapter establishes the fact that the growth in the e-commerce sector and extensive usage of online platforms and applications, for conducting almost every possible business operation, including online shopping has raised the risks of data breach and infringement of online operations (Reyes-Menendez *et al*.2019). This growing trend in the online shopping field had made people engage in sending reviews on particular services along with sharing their views, through these reviews.

However, critics find that in a considerable number of cases, these reviews are mal-manufactured, or essentially forged to show that the number of shares or likes that are shared on the particular product or service are fake and forged (Zhou *et al.*2020). This impacts the operations of the specific business organization in a significant way, so much so that it also develops the possibilities of recognizing spam messages and reviews in this regard. Just as positive reviews have a positive impact on the organization, negative reviews, also have a negative impact on the organization's operations (Paschen, 2019).

Hence, the discoveries of these illegal issues have become an important issue for providing relevant and useful information. Therefore, the need to detect the inaccurate updates and reviews on a particular product and service, is significantly important, as it helps in segregating between the accurate and original one, with the inaccurate and fake one (Gao *et al.*2019). Hence, this is done through the use of machine learning and the importance of doing such a detection is necessary to ensure that the company, which is producing and delivering the particular product or service to the target market does not develop any forged review from the audience (Molina *et al*.2021).

It is often seen that most of the online shoppers engage in sharing and posting reviews for any particular product or service, before participating in the actual purchase decision and process. This is where fake posts tend to have a strong impact on the minds of these buyers. Hence, it is important to detect these fake reviews and posts, so that the business organization does not get fluked from the fake reviews and come up with original views and opinions from the buyers and can engage in the right decision-making process, accordingly (Kaur *et al*.2020).

Hence, the primary aim is to ensure that the views shared on the online portal are genuine and does not have any fraudulent activity, engaged with it. The use of the machine learning technique for the current research would be based on using the sentiment analysis framework, wherein the unfair and forged reviews shall be detected, through the supervised learning process (Ruchansky *et al*.2017). The new system labelled as the Fraud Review Discovery shall be created to ensure that the technical team is able to detect the fake reviews and self-promoted views, and process the removal of the same.

The use of the sentiment classification technique within the machine learning algorithm of the Fraud Review Detection shall competently establish the need to detect fake reviews, for better business operations and ensure an accurate assessment of the views of the target audience, so that the decisions, to be considered by the organizations, depending on these reviews shall be appropriate (Martínez-Martínez *et al*.2017).

2.4. The selection of the behavioural approach along with the textual approach

Sinha *et al.* 2018 have explained that it is important for the customers to decide to buy any products and the behavioural approach of the customers is included. The customers have read the reviews that are provided by the other customers and the declarative memory of the customers. Cognitive behaviour is important as it helps in understanding whether the product that is sold is fake or fraudulent. These are included within the monetary gain; online reviews of the websites have become important areas where the spams and take products can be sold. Therefore, the authors have included that the customers have to judge the fairness of the product through proper judgment and proper constructive feedback had to be taken properly. Danish *et al.* 2019 have included that fake review detection can be done with proper considerable attendance. Therefore, the companies have started incorporating Yelp's algorithm to look into the detection of fake products and this helps in the

understanding of the filtered reviews that help in the manifestation of the product quantification, the algorithm helps in the understanding of review of the hosting sites. The algorithm helps in the development of the elimination of fake reviews and provides comprehensive coverage of the products included. The reviews are done with the behavioural approach that includes two kinds of phenomenology. First, supervised and unmonitored learning is introduced within the machine and then the linguistics and behavioural characteristics are shown according to the requirements. It is seen that the supervised learning scarcity of the machine has helped in the review of the products by 70% accuracy. The internet has helped in understanding the behavioural context of machine learning that has helped in understanding the quantity and online reviews that helped in reviews and comments in a more convenient manner.

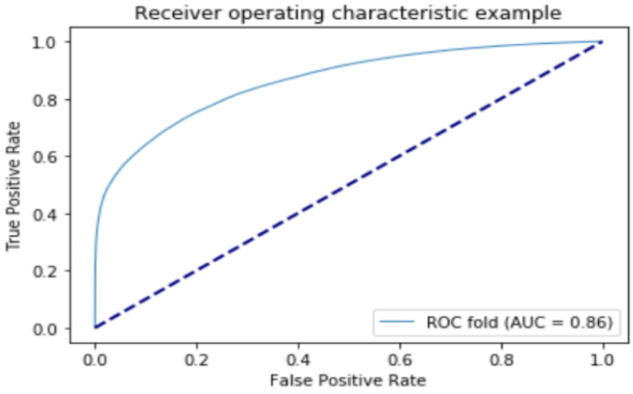

Figure 2: Understanding the false positive rate by Yelp's algorithm

(Source: Anon, (2021). Fake Product Review Detection using Machine Learning)

2.5. Significance of feature selection technique for the suggested approach

According to the concepts that are developed by Best *et al*, 2020 have shown that feature selection of the product reviews is to be done with the understanding of the behaviour and logistic approaches of the customers. Machine learning has helped in understanding artificial intelligence and deep learning scenarios to understand future prediction and analysis. The objective of the detection is to remove fake reviews from the internet. The selection of the features includes the integrity of the products, easy-to-find reviews, and providing illustrations of purchase decisions that are included towards the products. The algorithm of selection features is hidden by all the companies so when the product reviews are disclosed to the internet and view of protecting the interest of the customers and e-commerce portals. It is shown that Sentiment analysis has helped in providing supervised learning techniques that provide the understanding of the e-commerce portals. Yelp algorithm removes the fake product review by understanding the features of language and similar datasets that are made into the regressive analysis. Three different algorithms had and are included within linear regression algorithm, logical regression algorithm, and neural networks such as CNN and RNN networks that help in providing filtered supervised optimal accuracy rate of determination of the game reviews of the product. It is shown by the authors that supervised machine learning helps in understanding the features of the products and this is implied by the input systems of the algorithm and the sighting is favoured through the understanding if they integrate machine learning through neural connections. The review of the similarities of the dataset are considered and placed against the negative reviews and this is where the filtered and unfiltered data is shown. The supervised dataset is distinguished between the fair and unfair positive and negative reviews. This is done through the methods of collaborative filtering, removing the obstacles that help in understanding the optimal accuracy rate and review of the data sets.

Wahyuni and Djunaidy, 2016, have explained the core focus of the algorithm that helps in the understanding of the features that are differentiated. The features such as integrity products and collaborative filtering. The algorithm distinguishes the purchase decision and helps in providing the identification of the fake reviews. The feature engineering that is included in the length of the review, number of sentences, the average length of the reviewer, percentage of numerals, percentage of the capitalized words, and percentage of the positive and negative positions that are put in the words of the review. Deep learning will help in the understanding of these features. Each of the words review are learned with a deep simulation process and this helps in providing the governance that is required within the structural clarity of the words and the sentences

Wang *et al,* 2015 has explained that lexical features are fed to machine learning, and the artificial intelligence then functions according to the feed that is produced within the functionality of the algorithm. The algorithms are based on predictive analytics and this has helped in understanding the lexical attributes and the help in providing the content and style differentiation between the reviews. The spam detection methods are used that have been illustrated with the progressive understanding and the machine learning that lead to the understanding of the functionality but the sentences are made according to the features will help in providing the manifestation.

When the product reviews are disclosed to the internet and view of protecting the interest of the customers and e-commerce portals with proper manifestation that has been understood with the proper formulation of the features. It is shown that Sentiment analysis has helped in providing supervised learning techniques that provide the understanding of the e-commerce portals with proper manifestations of anagrams and semantic inconsistency. The authors have helped in the understanding of the supervised and unmonitored learning is introduced within the machine and then the linguistics and behavioural scarcity of the machine has helped in the review of the products by 70% accuracy.

2.6. Benefits of Fraud Review Discovery Summary

The research proposes the development of a Fraud Review Discovery system, which is powered by the use of the machine learning algorithm and shall allow the technical experts to administer and manage datasets and ensure removal of fake reviews on the online shopping web portals and e-commerce sites (Sharif *et al.*2019). This new proposed format shall help in determining the fact, whether the product or service review is true or has a false statement. The approach, for developing this new system of fake review detection and fraud review detection shall focus on obtaining results using Mechanical Learning Methods.

The primary focus of the approach is to emphasize on the aspects of the understanding and reviewing the content of the posts and views, shared by the users, in the form of the statement that are shared on the particular product or service (Yin *et al.*2017). By reviewing the statement reviews of the users, the Fraud Review Discovery system shall help in segregating the fake reviews, with the original reviews, by interpreting and understanding the choice of works and structure of changes. The use of the Neutral networks, including the CNN and RNN models would help in determining the focus and significance of running the Fraud Review Discovery system (Zhou *et al.*2015).

The Neutral networks, including the CNN and RNN models refer to the models that are used, within the framework of a resiliently developed convolutional neutral and systemized network, wherein a group of deep learning methods and networks shall be applied to analyze the visual imagery (Liu *et al.*2020). The use of the Neutral networks, including the CNN and RNN models is characterized by the fundamentals of the identifying the neutralized forms, in regards to identifying the neutral connection between the artificial neutral networks, followed by the recognition of the nodes that are developed based on the frameworks of accessing a directed graph , along the temporal issue (Kowsari *et al.*2017).

The CNN models are the types of neural network model which allows us to extract higher representations for the image content, followed by the recognition of the key differences between the two networks and accessing its neutrality (Li *et al.*2018). The major difference between the CNN and RNN networks, identify the ability to process the system of flowing the temporal information, allowing the transmission of data, in between the networks and understanding the data, found in the sentences, which help in understanding the sequences (Zulqarnain *et* al.2020). The series of sentence structures along with understanding the sequences of the words and arrangements of the views and reviews are detected.

With the use of the CNN and RNN features, and a neutralized networking approach, the technical experts are able to recognize the features, along with detecting the fake and true reviews, by using filters, which are also known as Kernels (Qin *et al.*2020). The approach that is used by the standardized aspects, of running a machine learning framework contributes to the identification of the true and fake words, wherein the filters work automatically and randomly (Paolanti *et al.*2018).

0.230	**0.380**	**0.971**
0.402	**0.119**	**0.886**
0.693	**0.563**	**0.771**

Figure 3: A filter, used in the CNN and RNN neutral network that helps in identifying the fake and true statements and reviews

(Source: Qin *et al.*2020)

By using the filters, it needs to be stated that the proposed method shall use the filters, in such a way so that the CNN network is able to identify the handwritten digits, by recognizing the characteristics, which includes the assessment of the enlarged version of a 28 x 28 pixel image, considering the MNIST dataset (Du *et al.*2017). The use of the CNN dataset helps in recognizing he ways and filters that shall be used to transform the data, by

conducting a sentimental classification, which identifies the difference between the true and fake reviews (Peng *et al.*2018).

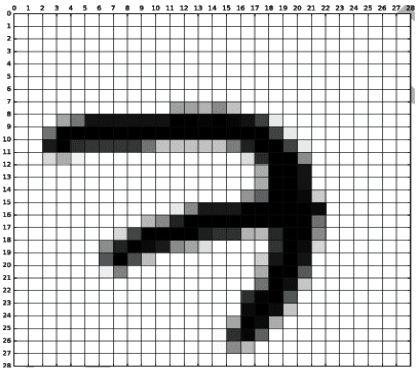

Figure 4: Image and letter detection before its conversion into pixel values

(Source: Du *et al.*2017)

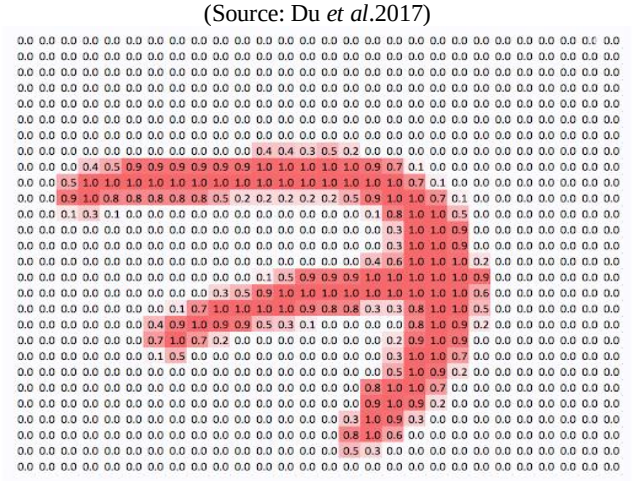

Figure 5: Image and letter detection, after its conversion to pixel values

(Source: Du *et al.*2017)

Hence, from the above image, it is evident that the use of machine learning, powered by the application of the CNN and RNN neutralized network, the use of filters plays an instrumental role in differentiating the fake and original reviews from the consumers on the online platform.

2.7 Literature Gap

An online product review of the shopping experience on social media encouraged the user to provide feedback. Today, many e-commerce sites allow a customer to write a review or comment on a product they purchased from that site. Customer reviews can build a good brand name or make a product famous. For this reason, in one product a customer reviews about a product have been submitted by an entity of the people themselves to act in order to provide a false review and one can degrade the product by giving a false negative review about the product. In this work, we will propose a framework for obtaining fake product reviews or spam reviews through Opinion Mining. The opinion mine is also known as Sentimental Analysis. In analyzing Sentiment, we try to get the customer's opinion on a piece of text.

We first take a review and evaluate whether the review is positive or negative or neutral using sensitive analysis. We use a spam dictionary to identify spam words in updates. In Text Mining we will use the Logistic regression, Decision tree classifier and the Neural network and compare their effect to what is most accurate and on the basis of these behaviors we obtain certain results

2.8 Summary

The chapter fundamentally introduces the significance of machine learning in today's business environment, along with justifying the use of Machine learning, as an effective tool in detecting fake reviews from online shoppers and facilitating a fair and accurate business operation. The model suggested for the development of the Fraud Review Discovery is the use of Neutral networks, including the CNN and RNN models. The chapter also studies the selection of the behavioral approach of detecting fake reviews, over the textual approach. Finally, the chapter discovers the benefits of the Fraud Review Discovery system, by explaining the approach, which is used to study the content of the consumer review text and the selection of the word choice.

3. Research Methodology

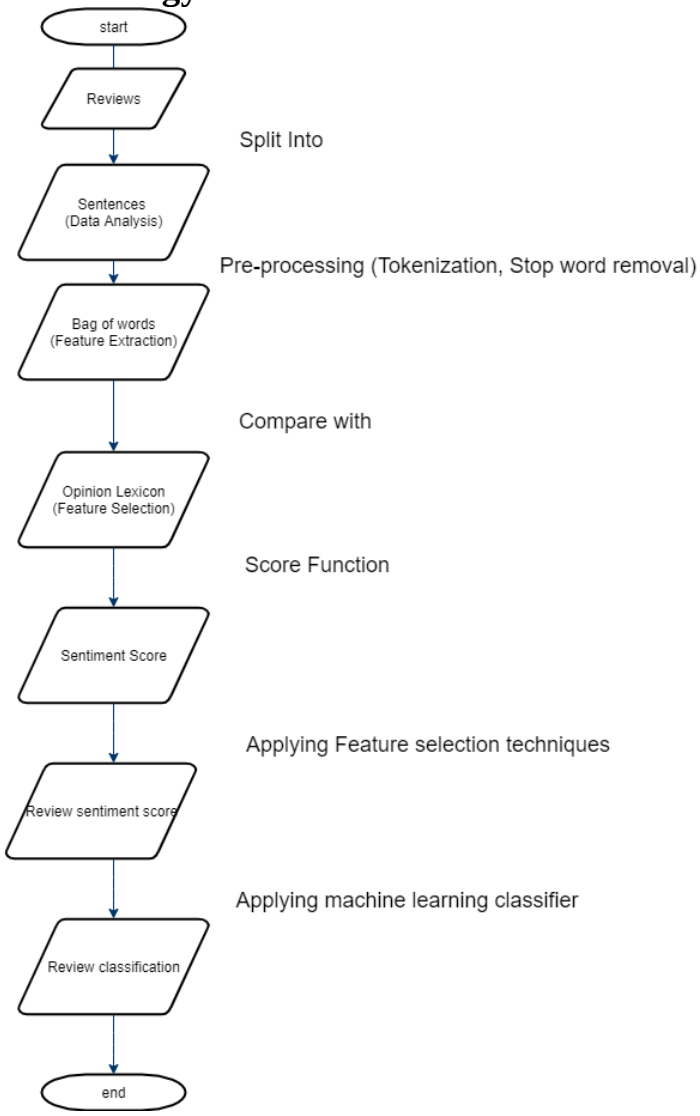

Figure 6: Research workflow Diagram

(Source: Self-created)

The figure above represents the research workflow that is followed for the research analysis completion. The workflow consists of scrapping the data in the form of product reviews from the online shopping website called Flipkart. The reviews are scrapped using the beautiful soup API using the python language. The reviews are gathered and collected in the comma separated file (CSV). Sentences, words and characters are reviewed from the dataset. Data features are extracted in the form of bags of words and data feature selection is carried out using the opinion lexicon analysis. The score function is defined to compare and score the sentiment of the product reviews provided by the product users. Product reviews ratings are analyzed with the application of feature selection techniques. Finally, the reviews are classified using the machine learning classifiers which include supervised learning approach, regression approach and the deep learning approach using the neural network

3.1. Data Summary

1. product_id: This data attribute describes the product identification number (ID)
2. product_title: This data attribute describes the product title displayed on the flipkart website
3. rating: This data attributes describes the rating of the product out of 5
4. summary: This data attributes is the summarization and description for the product.
5. Review: This data attribute consists of the review regarding the product
6. Location: This data attribute describes the location of the user or the reviewer
7. Date: This data attribute is the date of the review posted on the website
8. Upvotes: This attribute consists of the upvote (Positive) attribution of the product
9. Downvotes: This attribute consists of the downvote (Negative) attribution of the product

3.1.1. Data Pre-processing

```
# iterating over 50 pages of reviews
for i in tqdm(range(1,50)):

    URL = f"https://www.flipkart.com/apple-iphone-12-white-64-gb/p/itm8b88bdc03cd79?pid=MOBFWBYZTK33MBG9&lid=LSTMOBFWBYZ
    r = requests.get(URL)
    soup = BeautifulSoup(r.content, 'html.parser')

    cols = soup.find_all('div',attrs={'class':'col _2wzgFH'})

    for col in cols:
        row = col.find_all('div',attrs={'class':'row'})

        rating = row[0].find('div').text
        review = row[1].find('div').text

        dataset.append({'review': review , 'rating' : rating})
len(dataset)
```

```
100%|██████████████████████████████████████████████████████████████████████████| 49/49 [00:37<00:00,  1.32it/s]
0
```

Figure 7: Data Scrapping from the Flipkart website for product review purposes

(Source: Python code)

The figure depicts the data scrapping using the beautiful soup module API in python 3.8. The data is collected in the form of the product review from the web page of the online shopping site called Flipkart.

```
# Extracting all review blocks

row = soup.find_all('div',attrs={'class':'col _2wzgFH K0kLPL'})
```

```
# list to store data
dataset = []

# iteration over all blocks
for i in row:

    # finding all rows within the block
    sub_row = i.find_all('div',attrs={'class':'row'})

    # extracting text from 1st and 2nd row
    rating = sub_row[0].find('div').text
    review = sub_row[1].find('div').text

    # appending to data
    dataset.append({'review': review , 'rating' : rating})
```

Figure 8: Reviews Extraction for product 1

(Source: Python code)

	review	rating
0	It was nice produt. I like it's design a lot. ...	5
1	awesome sound....very pretty to see this nd th...	5
2	awesome sound quality. pros 7-8 hrs of battery...	4
3	I think it is such a good product not only as ...	5
4	This product sound is clear and excellent bass...	4
...	...	...
2865	GoodREAD MORE	5
2866	SuperbREAD MORE	5
2867	Nice...sound quality awsomeREAD MORE	5
2868	it was really nice and good one , i like this...	5
2869	Goog productREAD MORE	4

2870 rows × 2 columns

Figure 9: Review data frame for product 1

(Source: Python code)

```
# Extracting all review blocks

row = soup.find_all('div',attrs={'class':'col _2wzgFH'})
```

```
# list to store data
dataset = []

# iteration over all blocks
for i in row:

    # finding all rows within the block
    sub_row = i.find_all('div',attrs={'class':'row'})

    # extracting text from 1st and 2nd row
    rating = sub_row[0].find('div').text
    review = sub_row[1].find('div').text

    # appending to data
    dataset.append({'review': review , 'rating' : rating})
```

Figure 10: Reviews Extraction for product 2

(Source: Python code)

	review	rating
0	Delightful phone, the phone is just a peice of...	5
1	Excellent product worth for every penny, writi...	5
2	iPhone 6S Plus 64GB -> iPhone 12 128GBMy 2nd i...	5
3	The best is yet to come, I am really happy wit...	5
4	Night mode is simply amazing and give you a cl...	5
5	It's my first iPhone ever and I bought it with...	5
6	Green colour is charming and priceless No w...	5
7	The Product is fantastic with great nay awesom...	5
8	Bought First Apple product, Awesome design and...	4
9	Best ever delivery by flipkart, got this phone...	5

Figure 11: Review data frame for product 2

(Source: Python code)

```
dataset = []

# iteration over all blocks
for i in row:

    # finding all rows within the block
    sub_row = i.find_all('div',attrs={'class':'row'})

    # extracting text from 1st and 2nd row
    rating = sub_row[0].find('div').text
    review = sub_row[1].find('div').text

    # appending to data
    dataset.append({'review': review , 'rating' : rating})

dataset[:5]
```

```
[{'review': 'Price as per othe brand blue star is very high
  'rating': '5'},
 {'review': "Let me put all the doubts in place.1. Room siz
  'rating': '5'},
 {'review': "Review after 10 days of usage: I am happy with
  'rating': '5'},
 {'review': 'Its a super silent compressor which provides e
  'rating': '5'},
 {'review': "A very powerful ac from very powerful brand I
  'rating': '5'}]
```

Figure 12: Reviews Extraction for product 3

(Source: Python code)

	review	rating
0	Price as per othe brand blue star is very high...	5
1	Let me put all the doubts in place.1. Room siz...	5
2	Review after 10 days of usage: I am happy with...	5
3	Its a super silent compressor which provides e...	5
4	A very powerful ac from very powerful brand I ...	5
5	Very good product.. Cooling is very good.1.2 t...	5
6	Blue Star is the best AC brand in India. This ...	5
7	I am Refrigration Technician I like Blue Star ...	5
8	Great Product. Using this AC for almost a mont...	5
9	Me & my kid's r very much happy with this prod...	5

Figure 13: Review data frame for product 3

(Source: Python code)

```
SAMPLE_URL = "https://www.flipkart.com/boat-rockerz-400-bluetooth-headset/product-reviews/itm14d0416b87d55?pid=ACCEJ
r = requests.get(SAMPLE_URL)
soup = BeautifulSoup(r.content, 'html.parser')
print(soup.prettify()[:500])
```

```
<!DOCTYPE html>
<html lang="en">
 <head>
  <link href="https://rukminim1.flixcart.com" rel="preconnect"/>
  <link href="//static-assets-web.flixcart.com/www/linchpin/fk-cp-zion/css/app.chunk.8a1772.css" rel="stylesheet"/>
  <meta content="text/html; charset=utf-8" http-equiv="Content-type"/>
  <meta content="IE=Edge" http-equiv="X-UA-Compatible"/>
  <meta content="102988293558" property="fb:page_id"/>
  <meta content="658873552,624500995,100000233612389" property="fb:admins"/>
  <meta content="n
```

Figure 14: HTML component scrapping from the website's webpage

(Source: Python code)

```
from prettytable import PrettyTable
x = PrettyTable()
x.field_names = ["# Products", "# Reviews Per Page", "# Pages", "# Total Reviews Count"]
x.add_row([len(product_urls), 10, REVIEW_PAGES_TO_SCRAPE_FROM_PER_PRODUCT, len(product_urls)
print(x)
```

```
+------------+--------------------+---------+-----------------------+
| # Products | # Reviews Per Page | # Pages | # Total Reviews Count |
+------------+--------------------+---------+-----------------------+
|     10     |         10         |   100   |         10000         |
+------------+--------------------+---------+-----------------------+
```

Figure 15: Review analysis table format

Source: Python code)

```
# Readind the CSV file into data frame
df=pd.read_csv('/content/flipkart_reviews_dataset.csv')
df.head()
```

	product_id	product_title	rating	summary
0	ACCG2K38YCACC3XV	OnePlus Bullets Wireless Z Bass Edition Blueto...	5	Highly recommended
1	ACCG2K38YCACC3XV	OnePlus Bullets Wireless Z Bass Edition Blueto...	5	Great product
2	ACCG2K38YCACC3XV	OnePlus Bullets Wireless Z Bass Edition Blueto...	5	Excellent
3	ACCG2K38YCACC3XV	OnePlus Bullets Wireless Z Bass Edition Blueto...	5	Super!
4	ACCG2K38YCACC3XV	OnePlus Bullets Wireless Z Bass Edition Blueto...	5	Super!

Figure 16: Loading the main dataset into the data frame

(Source: Python code)

The figure depicts the final data set which is loaded into the data frame called df for further moel training, testing and evaluation.

```
df.columns
```

```
Index(['product_id', 'product_title', 'rating', 'summary', 'review',
       'location', 'date', 'upvotes', 'downvotes'],
      dtype='object')
```

Figure 17: Data Column analysis

(Source: Python code)

The above figure depicts the column of the data frame which consists of the column names namely product ID, product title, rating, summary, review, location, date, upvotes and downvotes.

```
Rows     :  7469
Columns  :  9

Features :  ['product_id', 'product_title', 'rating', 'summary', 'review', 'location', 'date', 'upvotes', 'downvotes']

Missing values :   3980

Unique values :
 product_id         10
```

(Source: Python code)

```
Unique values :
 product_id         10
product_title       8
rating              5
summary            85
review           4559
location         1024
date               63
upvotes           170
downvotes         102
dtype: int64
```

Figure 18: Unique data attributes analysis

(Source: Python code)

```
# Checking for null values in the data frame
df.info()
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7469 entries, 0 to 7468
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   product_id     7469 non-null   object
 1   product_title  7469 non-null   object
 2   rating         7469 non-null   int64
 3   summary        7469 non-null   object
 4   review         7469 non-null   object
 5   location       3489 non-null   object
 6   date           7469 non-null   object
 7   upvotes        7469 non-null   int64
 8   downvotes      7469 non-null   int64
dtypes: int64(3), object(6)
memory usage: 525.3+ KB
product_id          0
product_title       0
rating              0
summary             0
review              0
location         3980
date                0
upvotes             0
downvotes           0
dtype: int64
```

Figure 19: Null value analysis

(Source: Python code)

Using the n-gram tfidf vectorizer

```
[ ]  word_vectorizer = TfidfVectorizer(
         sublinear_tf=True,
         strip_accents='unicode',
         analyzer='word',
         token_pattern=r'\w{1,}',
         stop_words='english',
         ngram_range=(1, 3) )  # try 1,3
     #     max_features=10000)
     word_vectorizer.fit(all_text)
     train_word_features = word_vectorizer.transform(all_text)
```

Figure 20: Word Vectorization for the reviews

(Source: Python code)

```
char_vectorizer = TfidfVectorizer(
    sublinear_tf=True,
    strip_accents='unicode',
    analyzer='char',
    stop_words='english',
    ngram_range=(2, 6),
    max_features=50000)
char_vectorizer.fit(all_text)
train_char_features = char_vectorizer.transform(all_text)

train_features = hstack([train_char_features, train_word_features])
```

Figure 21: Character Vectorization for the reviews

(Source: Python code)

```
import time   # time1 = time.time(); time2 = time.time(); time_taken = time2 - time1
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
seed = 71

X_train, X_test, y_train, y_test = train_test_split(train_features, y, test_size=0.3,
print('X_train', X_train.shape)
print('y_train', y_train.shape)
print('X_test', X_test.shape)
print('y_test', y_test.shape)
```

```
X_train (5228, 119918)
y_train (5228,)
X_test (2241, 119918)
y_test (2241,)
```

Figure 22: Data Splitting into train and test data subset

(Source: Python code)

3.1.2. Feature Selection and Extraction

The feature extraction and selection process help in training the machine learning classifier using the Bag-of-words model as the base model. Further the results from this classifier will be utilized to predict the result using the testing data.

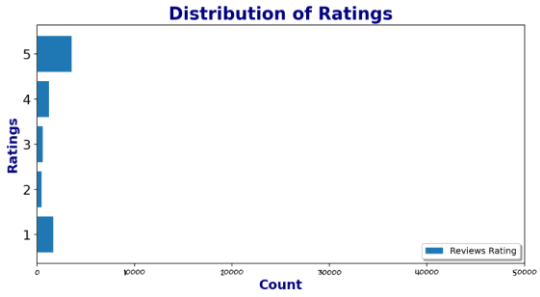

Figure 23: Product ratings analysis

(Source: Python code)

Figure 24: Word cloud for the product reviews

(Source: Python code)

We have studied the previous work conducted related to this topic and tried to analyze the thesis, the process of detecting spam, identifying a drawback in existing methods, preparing strategies and a solution for continuity or expansion in order to achieve detailed research analysis.

```
#what are the popular location
df['location'].value_counts()
```

```
Bengaluru          139
New Delhi          114
Hyderabad          103
Chennai             92
Mumbai              81
                  ...
Mangaldoi            1
Valsad               1
Ramanathapuram       1
Kalaburgi            1
Ujjain District      1
Name: location, Length: 1024, dtype: int64
```

Figure 25: Analysing the popular location of the review sources

(Source: Python code)

In this step, we combine the data from the data sets of multiple review sources into a single data frame.

```
# How is the summary for the product
df['summary'].value_counts()
```

```
Wonderful                              322
Must buy!                              237
Best in the market!                    220
Classy product                         218
Just wow!                              212
                                      ...
No good bass                             1
Ear pain                                 1
Wonderful product. Value for money.      1
WOW                                      1
Value for Money                          1
Name: summary, Length: 85, dtype: int64
```

Figure 26: Analysing the product summary

(Source: Python code)

```
# which are the popular products title?
df['product_title'].value_counts()
```

```
OnePlus Bullets Wireless Z Bass Edition Bluetooth Headset
Boult Audio ProBass Curve Neckband Bluetooth Headset
boAt Rockerz 235v2 with ASAP charging Version 5.0 Bluetooth Headset
boAt Airdopes 131 Bluetooth Headset
CatBull In-ear Bluetooth Headset
N2B MAGNET Red & k1 Pack of 2 Bluetooth Bluetooth Headset
Allmusic powerful driven bass with dynamic beats powered wireless Sports bluetooth
EDIO Bluetooth Headphone with Mic Earphone Bluetooth Headset Bluetooth Headset
Name: product_title, dtype: int64
```

Figure 27: Analysing the popular product title

(Source: Python code)

The process of breaking a string of text into phrases, words, symbols, or other logical elements called tokens. The purpose of tokenization is to test words in a sentence. Tokenization relies heavily on simple heuristics to classify tokens by following a few steps:

A. Tokens or words separated by a white space, punctuation marks or punctuation marks.

B. White spacing or punctuation marks may or may not be included depending on the need.

C. All the letters within the composite letter unit are part of the token. Tokens can be made of all alphabet letters, alphanumeric letters or only numeric letters. Tokens themselves can be separators. For example, in multiple programming languages, identifiers can be set up and mathematical operators without white spaces. Although it seems that this will appear as a single word or symbol, grammar actually takes the mathematical operator (token) as a separator, so most tokens are grouped together, which can still be separated by a mathematical operator.

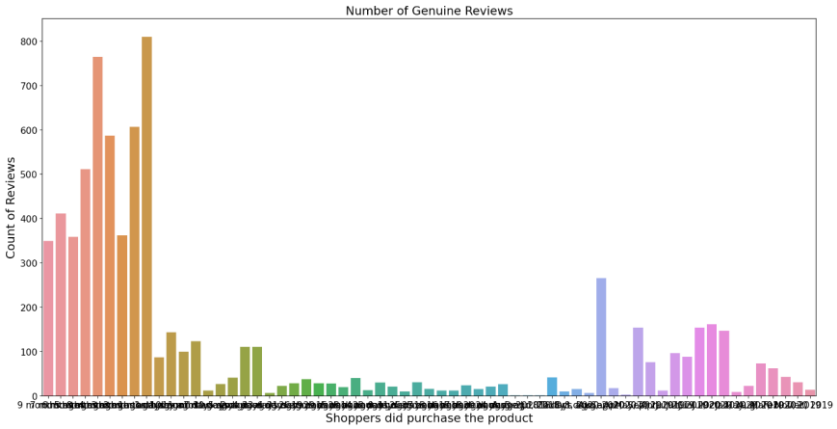

Figure 28: Analysing the genuine product reviews

(Source: Python code)

Common words will not help digging text mines such as prefixes, articles, and nouns can be considered default names. As all text texts refer to these words that are not required in the use of text mines. All of these words are deleted.

```
df['date'].value_counts()
```

```
1 month ago     810
8 months ago    765
2 months ago    607
4 months ago    587
5 months ago    511
                ...
23 days ago       7
Apr, 2020         3
Dec, 2018         1
Aug, 2018         1
Jul, 2018         1
Name: date, Length: 63, dtype: int64
```

Figure 29: Analysing the product review dates

(Source: Python code)

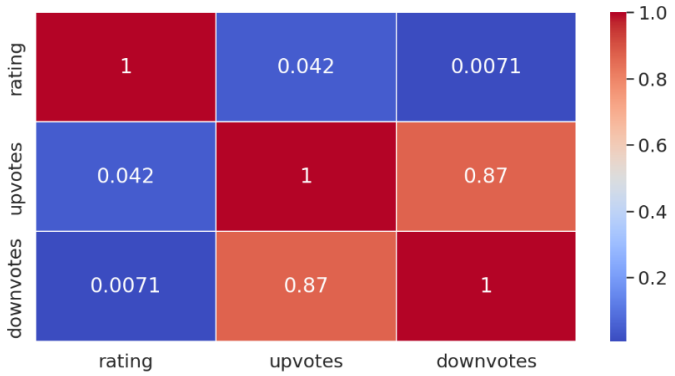

Figure 30: Correlation matrix for the numerical data attributes

(Source: Python code)

The Bag-of-word model is one of the simplest base models used in NLP. It makes a unigram model of text by keeping track of the number of occurrences of each word. This can later be used as a feature for Text Separators. In this word bag model, you only take individual words account and give each word a specific account of the account. We make a list of different words in the text corpus called vocabulary. After that, we can represent each sentence or document as a vector for each word represented as 1 at present and 0 for the absence of vocabulary. Another presentation could be counting the number of times each word appears in the text.

Which is the famous product?

```
[ ]  #df1 = df[df['reviews.did'] == True]
     df['product_id'].value_counts()
```

```
ACCFEZ99TZUFETH7    970
ACCFZGAPGBQ7FP8H    970
ACCG48F2SGVM72HN    919
ACCFS8HHZSEDBVDC    897
ACCFVA3KZVQWVTWG    897
ACCFVA3KZ2EYMYX3    885
ACCG2K38YCACC3XV    873
ACCFVPFPAEC3TEGP    600
ACCF5Y5ZDWXBEUQQ    232
ACCG3THXHSNNYRH9    226
Name: product_id, dtype: int64
```

Figure 31: Analysing the product ID

(Source: Python code)

(Source: Python code)

In this step, we use a variety of advance processing techniques to capture missing, noisy and inconsistent data. There are a number of pre-processing techniques such as erasing the folds of a folding dam, making word tokens, use of bag-of-word model, stop word elimination.

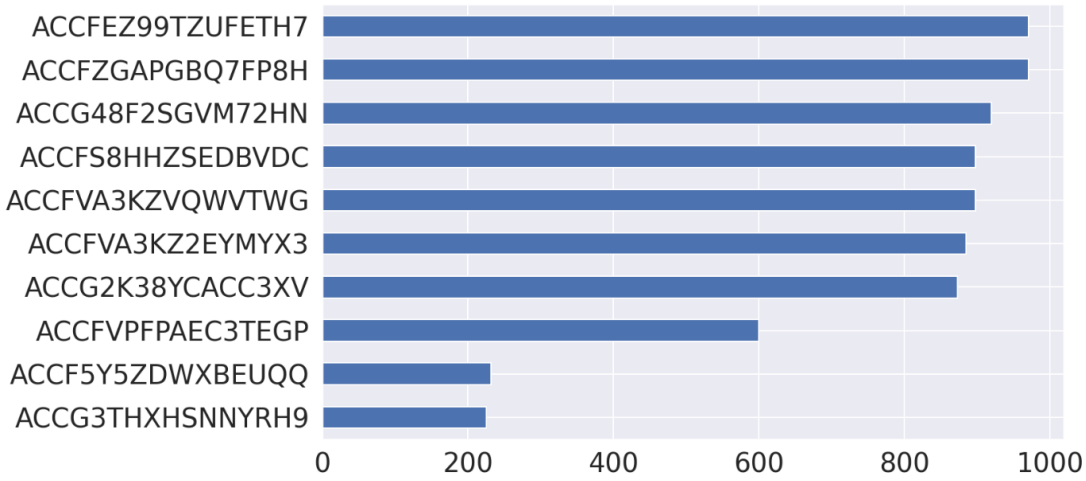

Figure 32: Analysing the popular product ID

(Source: Python code)

Common words will not help mining text mines such as prefixes, articles, and nouns can be considered default names. As all text texts refer to these words that are not required in the use of text mines. All of these words are eliminated after the data pre-processing.

4. Design Specification and Implementation

4.1. Model Evaluation

4.1.1. Logistic Regression

Since this is a classification problem, the use of logistic regression looks like a good start. Since the number of features at this time (over 4000) is very large the use of standard regularisation technique called LASSO to assist in selecting features for the workflow is necessary. The use of LASSO regularisation function aids in logistic regression model building is lesser training time. After the detailed analysis of the results obtained from the model summary it was observed that the model emphasised more on the verification purchase of each review for the product review prediction. One of the biggest disadvantages of using a regressive model for the prediction and classification task is that this model cannot be used with text mining techniques which would be better suited for the feature selection from the dataset which is in text format.

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import KFold, cross_val_score

time1 = time.time()

logit = LogisticRegression(C=1, multi_class='ovr')
logit.fit(X_train,y_train)
preds1 = logit.predict(X_test)

time_taken = time.time() - time1
print('Time Taken: {:.2f} seconds'.format(time_taken))
```

```
Time Taken: 12.31 seconds
```

Figure 33: Model training with Logistic regression algorithm

(Source: Python code)

4.1.2. Decision Trees

Decision tree is one of the supervised machine learning method, best suited for the classification purposes. This model yields better results with text-based features in classifying the product reviews. This advantage of using the decision tree over other supervised machine learning algorithms like random forest models is that it prevents the model from overfitting. With proper model parameter selection and model tuning accuracy score similar to logistic regression is achieved. The decision tree model can be considered to be superior to the previously used model because it yielded higher model performance along with capturing the underlying relationship with the text from the review dataset.

```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
```

```
from sklearn.tree import DecisionTreeClassifier
time1 = time.time()

classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train,y_train)
preds2 = classifier.predict(X_test)

time_taken = time.time() - time1
print('Time Taken: {:.2f} seconds'.format(time_taken))
```

```
Time Taken: 14.98 seconds
```

Figure 34: Model training with Decision tree algorithm

(Source: Python code)

4.1.3. Neural Networks

Neural network is selected as the deep learning approach for the fake review detection. The neural network consists of input layer, output layer, dropout layer, optimisation function and the regularization function. The neural network with 10 epoch size and 45 batch size yield the overall model accuracy of 90% for the testing data subset. The model also yielded the higher accuracy value of 97% for the training data subset. Each data columns are considered as the input in the neural network and the output layer consist of the classification result of whether the review is fake or not depending on the product review provided by the customers on the Flipkart website.

```
MAX_NB_WORDS = 20000

# get the raw text data
texts_train = train_text.astype(str)
texts_test = test_text.astype(str)

# finally, vectorize the text samples into a 2D integer tensor
tokenizer = Tokenizer(nb_words=MAX_NB_WORDS, char_level=False)
tokenizer.fit_on_texts(texts_train)
sequences = tokenizer.texts_to_sequences(texts_train)
sequences_test = tokenizer.texts_to_sequences(texts_test)

word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))
```

```
Found 5228 unique tokens.
```

Figure 35: Raw text vectorisation for the neural network model

(Source: Python code)

```
MAX_SEQUENCE_LENGTH = 200

# pad sequences with 0s
x_train = pad_sequences(sequences, maxlen=MAX_SEQUENCE_LENGTH)
x_test = pad_sequences(sequences_test, maxlen=MAX_SEQUENCE_LENGTH)
print('Shape of data tensor:', x_train.shape)
print('Shape of data test tensor:', x_test.shape)
```

```
Shape of data tensor: (5975, 200)
Shape of data test tensor: (1494, 200)
```

```
model = Sequential()
model.add(Embedding(MAX_NB_WORDS, 128))
model.add(LSTM(128, dropout=0.2, recurrent_dropout=0.2,input_shape=(1,)))
model.add(Dense(1, activation='sigmoid'))
```

```
model.compile(loss='binary_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])
```

```
model.fit(x_train, train_y,
          batch_size=128,
          epochs=10,
          validation_data=(x_test, test_y))
```

Figure 36: Model training with Neural network layers

(Source: Python code)

5. Evaluation

5.1. Logistic Regression model

	precision	recall	f1-score	support
1	0.92	0.73	0.81	630
2	0.15	0.96	0.26	24
3	0.14	0.88	0.24	26
4	0.21	0.66	0.32	131
5	0.97	0.69	0.81	1430
accuracy			0.71	2241
macro avg	0.48	0.79	0.49	2241
weighted avg	0.89	0.71	0.77	2241

Figure 37: Classification report for the logistic regression model

(Source: Python code)

```
[[459  95  47  20   9]
 [  0  23   1   0   0]
 [  1   0  23   1   1]
 [  1   4  20  87  19]
 [ 39  28  71 301 991]]
```

Figure 38: Confusion Matrix for logistic regression model

(Source: Python code)

PRODUCTS	PRECISION VALUE	RECALL VALUE	F1-SCORE
PRODUCT 1	0.92	0.73	0.81
PRODUCT 2	0.15	0.96	0.26
PRODUCT 3	0.14	0.88	0.24
PRODUCT 4	0.21	0.66	0.32
PRODUCT 5	0.97	0.69	0.81

5.2. Decision Tree model

	precision	recall	f1-score	support
1	0.82	0.76	0.79	535
2	0.42	0.58	0.49	108
3	0.34	0.38	0.36	143
4	0.43	0.58	0.49	301
5	0.85	0.75	0.79	1154
accuracy			0.70	2241
macro avg	0.57	0.61	0.58	2241
weighted avg	0.73	0.70	0.71	2241

Figure 39: Classification report for Decision tree model

(Source: Python code)

```
[[409  54  28  14  30]
 [ 26  63   5   4  10]
 [ 31  10  55  19  28]
 [ 12   5  20 174  90]
 [ 22  18  54 198 862]]
```

Figure 40: Confusion Matrix for Decision tree model

(Source: Python code)

PRODUCTS	PRECISION VALUE	RECALL VALUE	F1-SCORE
PRODUCT 1	0.82	0.76	0.79
PRODUCT 2	0.42	0.58	0.49
PRODUCT 3	0.34	0.58	0.36
PRODUCT 4	0.43	0.58	0.49
PRODUCT 5	0.85	0.75	0.71

5.3. Neural Network model

```
Epoch 1/10
47/47 [==============================] - 71s 1s/step - loss: 0.7292 - accuracy: 0.6631 - val_loss: 0.4163 - val_accuracy: 0.8561
Epoch 2/10
47/47 [==============================] - 73s 2s/step - loss: 0.3694 - accuracy: 0.8817 - val_loss: 0.3319 - val_accuracy: 0.8855
Epoch 3/10
47/47 [==============================] - 74s 2s/step - loss: 0.2489 - accuracy: 0.9183 - val_loss: 0.3596 - val_accuracy: 0.8574
Epoch 4/10
47/47 [==============================] - 73s 2s/step - loss: 0.2457 - accuracy: 0.9207 - val_loss: 0.3174 - val_accuracy: 0.8822
Epoch 5/10
47/47 [==============================] - 64s 1s/step - loss: 0.2001 - accuracy: 0.9324 - val_loss: 0.2848 - val_accuracy: 0.9070
Epoch 6/10
47/47 [==============================] - 62s 1s/step - loss: 0.1489 - accuracy: 0.9555 - val_loss: 0.2934 - val_accuracy: 0.9043
Epoch 7/10
47/47 [==============================] - 64s 1s/step - loss: 0.1296 - accuracy: 0.9600 - val_loss: 0.3017 - val_accuracy: 0.9076
Epoch 8/10
47/47 [==============================] - 66s 1s/step - loss: 0.1157 - accuracy: 0.9645 - val_loss: 0.2990 - val_accuracy: 0.9103
Epoch 9/10
47/47 [==============================] - 66s 1s/step - loss: 0.1091 - accuracy: 0.9675 - val_loss: 0.3082 - val_accuracy: 0.9103
Epoch 10/10
47/47 [==============================] - 66s 1s/step - loss: 0.1011 - accuracy: 0.9707 - val_loss: 0.3255 - val_accuracy: 0.9050
<keras.callbacks.History at 0x7f81e4c35b50>
```

Figure 41: Evaluation metrics for the neural network model

(Source: Python code)

MODEL	ACCURACY VALUE	TOTAL MODEL LOSS VALUE	VALANCE ACCURACY VALUE	VALANCE LOSS VALUE
NEURAL NETWORK WITH 10 EPOCH SIZE AND 47 BATCH SIZE	0.9707	0.1011	0.9050	0.325

5.4. Comparative Analysis

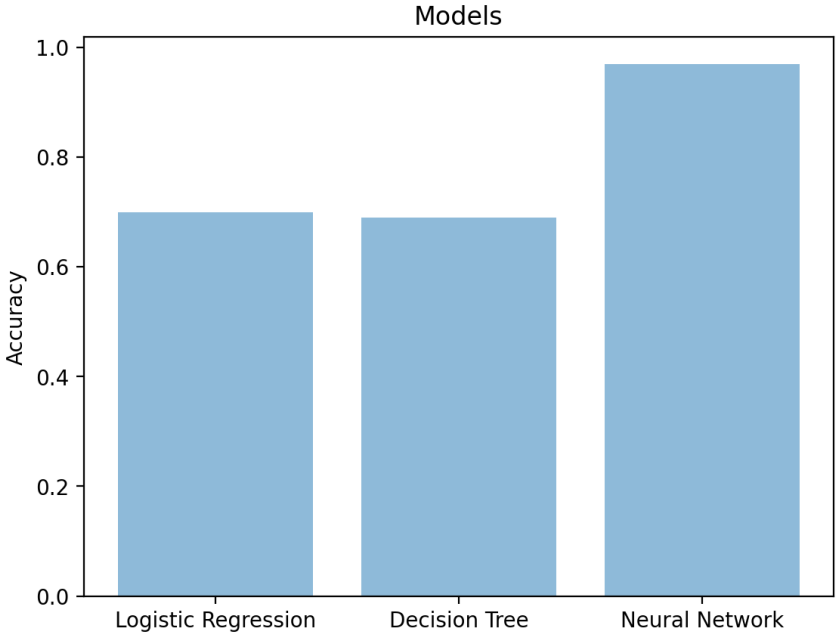

Figure 42: Comparative analysis for three models

(Source: Python code)

MODELS	ACCURACY VALUE
LOGISTIC REGRESSION MODEL	0.71
DECISION TREE MODEL	0.70
NEURAL NETWORK MODEL	0.97

From our work we have come to the conclusion that finding spam ideas in large amounts of unstructured data has become an important research problem. Although, some of the algorithms used in the spam analysis of ideas give good results, but still no algorithm can solve all the challenges and difficulties faced by today's generation. It is very important to consider specific quality standards such as usefulness, helpfulness and usability while analyzing each review. In literature research there are many complex explanations that describe the analysis of emotions in relation to various aspects. Our app that will help the user to pay for the right product without getting into any scams. Our work performed the analysis and map the genuine review to genuine product.
And the user can be sure about the product availability through the review prediction process. In the future we will try to improve the way we calculate sentimental feedback score. We will also try to update our data dictionary containing the sentiment words. We can try to add more words to our dictionary and revise the weights given to those words in order to get the most accurate counting points for updates. Sentimental analysis or opinion can be applied to any new classifier that follow the rules of data mining. Guide to future research is system utilization and performance evaluation using the proposed method for various measurement data sets. The main purpose of our work is to create a system that will receive spam and unwanted updates and filter them so that the user can understand the product information. The aim of our project is to improve customer satisfaction and make online shopping more reliable. The project will detect the fake reviews by incorporating mining algorithms like logistic regression classifier, Decision tree classifier and neural network classifier.

6. Error Analysis:

Error analysis helps to isolate, verify, and confirm erroneous ML estimates, thereby helping to understand the high and low performance of the model. The neural network gave the accuracy score of 97% but it varies in subgroup of the data, so the model performance changes if the input conditions are varied leading to the failure of the overall model performance.

7. Conclusion and Future work

In this study, the importance of reviews and how they affect almost everything related to web-based data was analyzed. It is clear that reviews play an important role in public judgment. Therefore, detecting fake reviews is clear and consistent in the research area.

In this paper, a machine learning machine that detects fake reviews is displayed. In the proposed approach, both the characteristics of the review and the behavioral characteristics of the reviewers are considered. The Flipkart dataset was used to evaluate the proposed approach. There are different classifications approach used like the supervised machine learning, logistic regression and the deep learning with neural network model. The developed approach uses and compares bi-gram and trigram language models. The results show that the neural network model performs better than the rest of the classifiers in the process of identifying fake reviews. Also, the results show that the behavior and textual analysis is considered critical features for the fake review prediction. Not all current work reviewers' models and techniques consider behavioral characteristics. Future work may consider adding features such as the frequency with which reviewers review and other behavioral characteristics over time.

Reviewers take reviews to complete and how often they submit positive or negative reviews. It is highly expected that improving the performance of the process of detecting fake reviews submitted by taking into account more behavioral characteristics.

References

[1] Bist, J., Hulsurkar, N., Bhalerao, S., and Narkhede, D., 2020. Comment Sentiment Analysis and Fake Product Review Detection.Available on:https://www.academia.edu/download/64555513/IRJET-V7I594.pdf

[2] Danish, N.M., Tanzeel, S.M., Usama, N., Muhammad, A., Martinez-Enriquez, A.M. and Muhammad, A., 2019, September. Intelligent interface for fake product review monitoring and removal. In 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE) (pp. 1-6). IEEE.Available on:https://ieeexplore.ieee.org/abstract/document/8884529/

[3] Sinha, A., Arora, N., Singh, S., Cheema, M., and Nazir, A., 2018. Fake product review monitoring using opinion mining. International Journal of Pure and Applied Mathematics, 119(12), pp.13203-13209.Available on:https://www.acadpubl.eu/hub/2018-119-12/articles/5/1203.pdf

[4] Wahyuni, E.D. and Djunaidy, A., 2016. Fake review detection from a product review using a modified method of iterative computation framework. In MATEC Web of conferences (Vol. 58, p. 03003). EDP Sciences.Available on:https://www.matec-conferences.org/articles/matecconf/abs/2016/21/matecconf_bisstech2016_03003/matecconf_bisstech2016_03003.html

[5] Boutaba, R., Salahuddin, M.A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F. and Caicedo, O.M., 2018. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, *9*(1), pp.1-99.

[6] Brunton, S.L., Noack, B.R. and Koumoutsakos, P., 2020. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, *52*, pp.477-508.

[7] Cavalcante, I.M., Frazzon, E.M., Forcellini, F.A. and Ivanov, D., 2019. A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management*, *49*, pp.86-97.

[8] Choudhury, A.M. and Nur, K., 2019, January. A machine learning approach to identify potential customer based on purchase behavior. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 242-247). IEEE.

[9] Dargan, S., Kumar, M., Ayyagari, M.R. and Kumar, G., 2020. A survey of deep learning and its applications: a new paradigm to machine learning. *Archives of Computational Methods in Engineering*, *27*(4), pp.1071-1092.

[10] Dey, A., Rafi, R.Z., Parash, S.H., Arko, S.K. and Chakrabarty, A., 2018, June. Fake news pattern recognition using linguistic analysis. In *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 305-309). IEEE.

[11] Dogru, A.K. and Keskin, B.B., 2020. AI in operations management: applications, challenges and opportunities. *Journal of Data, Information and Management*, *2*(2), pp.67-74.

[12] Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.W., Han, Z. and Pham, B.T., 2020. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides*, *17*(3), pp.641-658.

[13] Du, J., Gui, L., Xu, R. and He, Y., 2017, November. A convolutional attention model for text classification. In *National CCF conference on natural language processing and Chinese computing* (pp. 183-195). Springer, Cham.

[14] Gao, C., Li, X., Zhou, F. and Mu, S., 2019. Face liveness detection based on the improved CNN with context and texture information. *Chinese Journal of Electronics*, *28*(6), pp.1092-1098.

[15] Hasan, M., Islam, M.M., Zarif, M.I.I. and Hashem, M.M.A., 2019. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet of Things*, *7*, p.100059.

[16] Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), pp.305-311.

[17] Kaur, S., Kumar, P. and Kumaraguru, P., 2020. Automating fake news detection system using multi-level voting model. *Soft Computing*, *24*(12), pp.9049-9069.

[18] Kowsari, K., Brown, D.E., Heidarysafa, M., Meimandi, K.J., Gerber, M.S. and Barnes, L.E., 2017, December. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (pp. 364-371). IEEE.

[19] Kraus, M., Feuerriegel, S. and Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), pp.628-641.

[20] Leo, M., Sharma, S. and Maddulety, K., 2019. Machine learning in banking risk management: A literature review. *Risks*, *7*(1), p.29.

[21] Li, Y., Wang, X. and Xu, P., 2018. Chinese text classification model based on deep learning. *Future Internet*, *10*(11), p.113.

[22] Liu, Z., Huang, H., Lu, C. and Lyu, S., 2020. Multichannel cnn with attention for text classification. *arXiv preprint arXiv:2006.16174*.

[23] Martínez-Martínez, F., Rupérez-Moreno, M.J., Martínez-Sober, M., Solves-Llorens, J.A., Lorente, D., Serrano-López, A.J., Martínez-Sanchis, S., Monserrat, C. and Martín-Guerrero, J.D., 2017. A finite element-based machine learning approach for modeling the mechanical behavior of the breast tissues under compression in real-time. *Computers in biology and medicine*, *90*, pp.116-124.

[24] Molina, M.D., Sundar, S.S., Le, T. and Lee, D., 2021. "Fake news" is not simply false information: a concept explication and taxonomy of online content. *American behavioral scientist*, *65*(2), pp.180-212.

[25] Mullainathan, S. and Spiess, J., 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), pp.87-106.

[26] Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., Frontoni, E. and Loncarski, J., 2018, July. Machine learning approach for predictive maintenance in industry 4.0. In *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)* (pp. 1-6). IEEE.

[27] Paschen, J., 2019. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*.

[28] Peng, H., Li, J., He, Y., Liu, Y., Bao, M., Wang, L., Song, Y. and Yang, Q., 2018, April. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 world wide web conference* (pp. 1063-1072).

[29] Qin, Q., Hu, W. and Liu, B., 2020, July. Feature projection for improved text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8161-8171).

[30] Reyes-Menendez, A., Saura, J.R. and Filipe, F., 2019. The importance of behavioral data to identify online fake reviews for tourism businesses: A systematic review. *PeerJ Computer Science*, *5*, p.e219.

[31] Ruchansky, N., Seo, S. and Liu, Y., 2017, November. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797-806).

[32] Shang, C. and You, F., 2019. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. *Engineering*, *5*(6), pp.1010-1016.