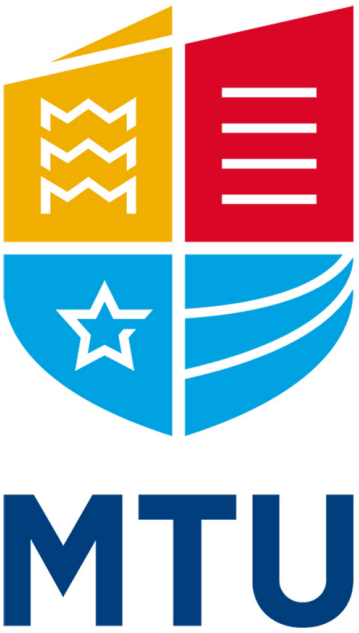

Detecting malicious reviews on online E-Commerce platforms

by

Saurabh Gaglani

This thesis has been submitted in partial fulfillment for the degree of Bachelor of Science in Software Development

in the
Faculty of Engineering and Science
Department of Computer Science

May 2023

Declaration of Authorship

This report, Detecting malicious reviews on online E-Commerce platforms, is submitted in partial fulfillment of the requirements of Bachelor of Science in Software Development at Munster Technological University Cork. I, Saurabh Gaglani, declare that this thesis titled, Detecting malicious reviews on online E-Commerce platforms and the work represents substantially the result of my own work except where explicitly indicated in the text. This report may be freely copied and distributed provided the source is explicitly acknowledged. I confirm that:

- This work was done wholly or mainly while in candidature Bachelor of Science in Software Development at Munster Technological University Cork.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Munster Technological University Cork or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Saurabh Mahesh Gaglani

Date: 22/12/2022

MUNSTER TECHNOLOGICAL UNIVERSITY CORK

Abstract

Faculty of Engineering and Science
Department of Computer Science

Bachelor of Science

by Saurabh Gaglani

Online content rating systems allow users to locate new content by leveraging the rating of other content. The impact of these systems on small and large businesses alike has grown remarkably over the last few years. Yelp alone claims to have over 183 million reviews worldwide. However, online reviews becoming one of the crucial determinants of a business' success or failure has given rise to many bad actors who use unethical means to improve their reputation by 'review-bombing' competitors with fake reviews. In this paper, we will look to use text analysis, sentiment analysis and supervised learning to develop an algorithm that can identify and flag malicious reviews. The contributions of this paper will be to (i) Compare two popular datasets to train a machine learning model (ii) Create a supervised learning model for fake review detection (iii) Test whether an unsupervised learning model for fake review detection can be created using topic modelling.

Acknowledgements

Thank you to my supervisor Alex Vakaloudis without whose help and guidance this research would not have been possible....

Contents

List of Figures

List of Tables

Abbreviations

LAH **L**ist **A**bbreviations **H**ere

For/Dedicated to/To my...

Chapter 1

Introduction

Online reviews have become a very important factor for a consumer to find and share content and have become an important information source that facilitates consumers to make purchase decisions. Studies have shown a very high correlation between the decision to buy a product and the number of positive reviews it gets[1]. It is clear that the reviews a business gets directly impacts its sales and revenue and unfortunately, this increasing influence of online reviews have made them an attractive target for manipulation. Increasing a business's overall rating on Yelp by just one star can lead to 9 percent increase in revenue [2], which means there is potential for real financial gain in manipulating reviews. This gives rise to bad actors that manipulate reviews for their own personal gain. These malicious users manipulate reviews in many different ways. For example because online retailers like Amazon and Yelp rely on individual sellers. Their platform removes sellers if they have any reason to suspect review manipulation on their part. The malicious users use this to their advantage and 'review-bomb' their competitors with many fake positive reviews so that they get banned from selling their wares on well known online platforms. They can also buy positive or negative ratings from otherwise legitimate users or from bots to either inflate their own rating with fake positive reviews or to flood a competitor with negative reviews. It is very important to online retailers and customers alike that the integrity of online reviews remain unquestionable. For this reason fake review detection has attracted significant attention from both businesses and the research community.

1.1 Aims

Since it was first studied in 2007[3], a lot of progress has been made in the area of fake review detection and online retailers use a lot of different ways to stop fake reviews from

being posted on their websites. The main detection technique used has been supervised learning while the area of unsupervised learning has been left largely unexplored. To train these supervised learning models, lots of companies create labelled datasets and some of these datasets have also been made available to the public. Two of the most popular of these is the Amazon review dataset and the Yelp Review dataset. Although some work has been done using these, their effectiveness for training supervised and unsupervised learning models has not been clearly laid out and a gold-standard dataset for fake review detection does not exist yet. The objectives of this paper are to -

1. Use a labelled dataset created from Yelp reviews that was used in the paper, **a framework for fake review detection in online consumer electronic retailers**[4] and train different supervised learning models using this dataset to verify the previous findings.

2. Use Natural Language Processing and Natural Language Understanding to create an unsupervised learning model using topic modelling and if successful, use the model to label the Amazon review dataset and separate reviews into 'Genuine' and 'Fake'.

3. Run the same supervised learning models on both the Amazon and the Yelp datasets to assess whether these datasets are suitable for training machine learning models, and which dataset is more suited for identifying fake reviews in e-commerce platforms.

1.2 Motivation

Review manipulation is still a huge problem for consumers and site operators in the online marketplace. Recently, there have been plenty of news stories about rising concerns over fake reviews – and businesses that suppress negative content. In fact, governments are even stepping in to squelch this growing problem. In 2019, the expanded it's regulations to protect consumers by including stricter demands on fake and misleading reviews. And in 2021, the Federal Trade Commision (FTC) in the United States fined clothing retailer Fashion Nova 4.2m dollars for suppressing negative reviews. To maintain a fair and balanced marketplace for all consumers, it is essential that we build review systems that cannot be tampered with and any researchers are working on doing just that. One of the most important factors of any Machine Learning related research is to find a good dataset and even though the Amazon and Yelp datasets are quite popular, there is no literature on whether these datasets are effective or not. I have decided to pick this

topic of research to fill this gap of gold-standard data and help the research community build machine learning models that are more effective in stopping malicious reviews.

1.3 Contribution

This research will require a very detailed understanding of data analysis in order to achieve it's objective. It will also require an understanding of Artificial intelligence. More specifically, text analysis in machine learning and the different methods of performing text analysis including but not limited to sentiment analysis, topic modelling, intent detection, topic analysis, concordance, collocation and clustering.

1.4 Structure of This Document

In Chapter 2 the guidance in structuring the literature review is given. Chapter 3 describes the main requirements for the problem definition, 4 discusses the implementation of the study and 7 provides the conclusions of the study.

Chapter 2

Background

In this chapter we will cover related work that has been done in fake review detection using Machine Learning. We will also discuss the current state of the art when it comes to dealing with fake reviews and vote buying.

2.1 Thematic Area within Computer Science

2.1.1 CCS Concepts

• Computing methodologies Artificial intelligence Natural language processing Information extraction

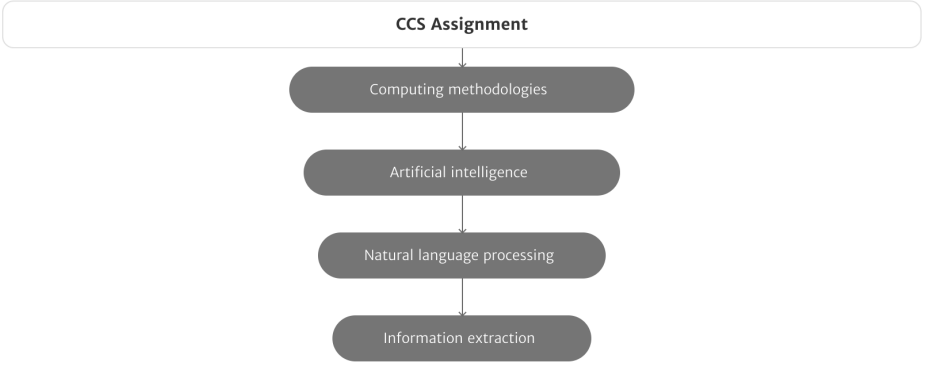

FIGURE 2.1: CCS classification

2.1.2 Keywords

Natural Language Processing, Sentiment Analysis, Supervised Learning, Unsupervised Learning, E-commerce

2.2 An Analysis of Malicious Review Detection Techniques, Tools and Algorithms

There have been various effective tools and techniques that have been developed over the years to detect fake reviews. Jindal and Liu proposed three basic approaches to identify fake reviews these are -

2.2.1 Approaches

- Review Centric Approach - This approach uses text analysis on the contents of the review to detect whether a review is fake or not. It checks various features such as sentence structure, number of positive and negative sentiment words in the review, review content similarity, use of capital letters etc[5].
- Reviewer Centric Approach - This approach considers the user that writes the review and features that are related to them. The features used in this approach are - age, profile picture, number of friends, number of reviews written, average review length, average review rating, IP address of the reviewer etc[5].
- Product Centric Approach - This method focuses on the product that the review is being written for and takes into account factors such as - Price of the product, number of sales, sales rank of the product, average rating etc[5].

The review centric approach has been proposed in many papers. In unsupervised learning, Wahyuni and Djunaidy created an Iterative Computational Framework to classify fake and genuine reviews and concluded that the use of ratings alone to assess whether a review is fake or genuine is inadequate and that by using text properties the precision value can be improved by 6%[6]. This gave rise to many researchers using classification algorithms to solve this problem.
In supervised learning, Ott et all[7]. Found that using the combination of psycholinguistically motivated features and bigram features can achieve 89.6% cross-validated accuracy on their novel dataset generated using Amazon Mechanical Turk. However, when Mukherjee et all.[8] attempted to use this method on a real life yelp dataset they

achieved only 67% accuracy. They extended on the method proposed by Ott et al. by adding behavioural features to the n-gram features thus combining the Review Centric and the Reviewer Centric approach. They also concluded that AMT fake reviews are not representative of real life fake reviews because AMT fake reviews are weak in detecting real fake reviews. Finally, Barbado et all.[4] verified the findings of [8] and concluded that a combination of the Review Centric and the Reviewer Centric Approach achieves the best accuracy when it comes to fake review detection on real life datasets.

Reference	Approach	Algorithm	Review Centric	Reviewer Centric	Product Centric
Barbado et al.[4]	Supervised	NB, DT, GNB	x	x	
Mukherjee et al.[8]	Supervised	SVM	x	x	
Ott et al.[7]	Supervised	SVM, NB	x		
Fusilier et al.[9]	Semi-Supervised	PU Learning	x		
Wahyuni and Djunaidy[6]	Unsupervised	ICF ++	x	x	x
Zhao et. al[10]	Supervised	CNN	x		
Li et al.[11]	Supervised	NB, DT, GNB	x		
Birim et al.[12]	Semi-Supervised	Topic Modelling	x		

TABLE 2.1: Related Work

2.2.2 Tools

There are now a few tools available online that employ these validation techniques and can be used for free by consumers that want to know whether a products reviews are genuine or not. Of the tools available online, two of the most notable are -

1. **Fakespot**
   Fakespot is a data analytics platform that strives to change individuals' perceptions of reviews. It is available to consumers as a free google chrome extension. Fakespot uses English language pattern recognition, the profile of the reviewer and correlation with other reviewer data. Their algorithm uses machine learning to constantly improve itself by looking at profile clusters, sentiment analysis and cluster correlation[13].

2. **Pasabi**
   Pasabi is a digital trust and safety platform that prevents multiple threats to online e-commerce websites using one integrated solution[15]. Pasabi uses natural Language Processing and supervised learning classifiers that are trained using hundreds of thousands of real world examples[16]. Along with machine learning, It also uses clustering to map out reviewers and the phone numbers linked to their account to flag a user if multiple accounts are created using the same phone number or email, Figure 2.3 shows an example of the cluster map created by Pasabi[14].

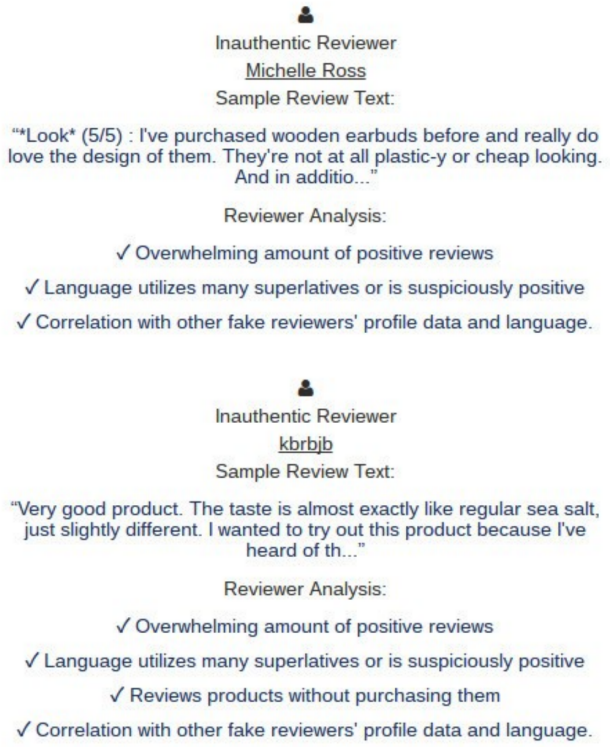

FIGURE 2.2: A flagged inauthentic review on Fakespot [13]

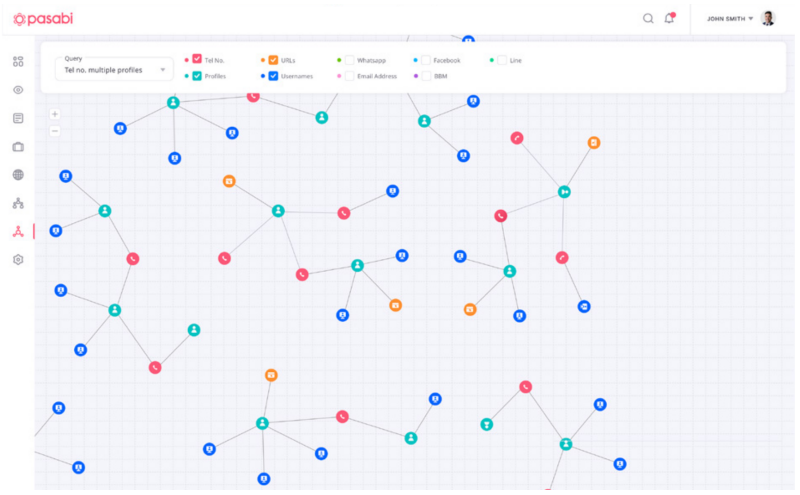

FIGURE 2.3: Cluster Map on Pasabi [14]

There aren't a lot of other reputable tools online that can detect fake reviews which shows the need for this research and more research to be conducted in this field in the future.

2.2.3 Techniques and Algorithms

The main contribution of this paper is to find which data set is more suited for fake review detection using Artificial Intelligence. A lot of websites do not collect user information such as number of prior reviews, profile picture etc. and it is also clear from Table 2.1 that the most successful algorithms are created by using reviews as a feature vector. For this reason, we will mainly focus on the review centric approach. Most of the prior work done in identifying fake reviews has used techniques that are part of Natural Language Processing and all these works have achieved some form of success, therefore performing NLP on the review dataset will be the best indicator on whether the dataset is suited to predict fake reviews or not. This subsection will discuss popular Natural Language processing techniques and provide some background on the techniques used in this paper.

2.2.3.1 Sentiment Analysis

Sentiment Analysis also known as Opinion Mining(OM) is a very popular Natural Language Processing technique used to determine whether the sentiment of some given data is positive negative or neutral, the sentiment of a sentence can be used to find patterns in fake reviews and identify them. Supervised learning algorithms like KNN, NB, SVM etc. can be used to perform sentiment analysis. To train these algorithms, a lot of the other research work uses the well known but simplistic method of a traditional Bag of Words (BOW). In the bag-of-word representation, only the meaning of the words is considered and the association between them are not represented. Outcomes when using this approach will be mostly mediocre, since BOW loses many subtleties of a possible good representation like consideration of word ordering and consideration of word context. To improve the results obtained by sentiment analysis, we must be able to map the syntactic and semantic relationship between the words into the feature vector, this is called Word Embedding. There are many readily available open sourced tools and techniques used for Word Embedding -

1. **Word2vec**
   Word2vec is a method developed by google to capture the context of words and reduce the dimensions of the feature vector by removing similar words from it. It does this by using a concise three-layer neural network structure[[15]. Word2vec has two different methods as shown in Figure 2.4, these are - Continuous Bag of Words(CBOW) and Skip-gram. In the CBOW method, the goal is to predict a word given a sliding window of the surrounding words(Figure 2.5) while in Skip-gram the model predicts the window of surrounding words given the current word.

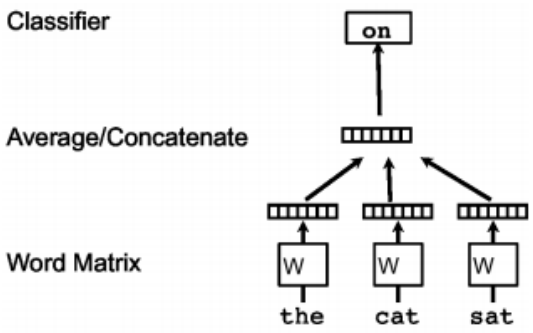

FIGURE 2.4: CBOW Method, uses the words "the", "cat" and "sat" to predict the word "on" [16]

2. **Doc2Vec**

   Doc2Vec is built on the CBOW method of Word2Vec. It uses the implementation of Continuous Bag of Words and adds a feature vector called "Paragraph Id" to it as shown in Figure 2.5, after training, the Paragraph Id contains a numerical representation of the document in it.

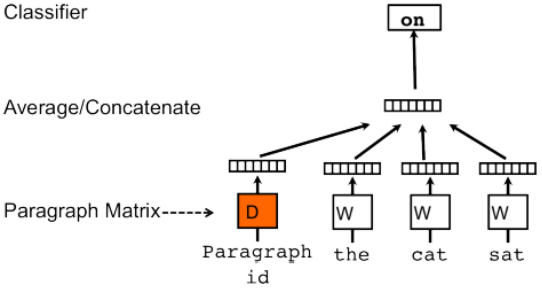

FIGURE 2.5: Doc2Vec adds a feature vector called "Paragraph Id" to the CBOW method. [16]

There are various Artificial Intelligence techniques that can be applied to the word vectors that have been trained by Word2Vec or Doc2Vec. Predicting whether a given review is real or fake is clearly a classification problem as there are two discrete class labels that we are looking to predict. Therefore we will look at some of the more widely used classification algorithms. Popular classification algorithms that have been used to predict fake reviews are -

Naive Bayes Classifier

Naive Bayes Classifiers are classifiers that operate on the assumption that all the features that predict the target value are independent of each other. They are probabilistic classifiers that can predict class membership probabilities, such as the probability that a given sample belongs to a particular class[17]. This shows that performing sentiment analysis using a Naive Bayes classifier is a credible method for predicting whether a

review is fake or not. The Gaussian Naive Bayes formula operates using the equation -

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Many previous works have done sentimental analysis using a Naive Bayes classifier to identify fake reviews[4][9][11] and achieved respectable results.

Support Vector Machines

Support Vector Machine(SVM) is a supervised learning model that can be used in both regression and classification. The SVM classifier is defined by the separating hyperplane i.e. This algorithm receives the labelled data and maps it into a 2 dimensional vector space, then it draws a seperating hyperplane, this hyperplane separates data points according to it's class and predicts the class of a new point by checking which side of the hyperplane it should fall under [18]. The SVM algorithm has also been used extensively in fake review detection [11][19].

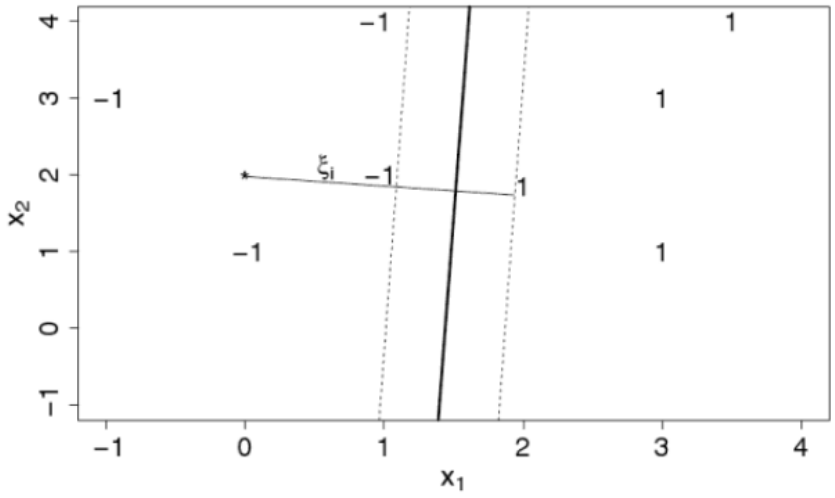

FIGURE 2.6: Soft Margin SVM in a 2D vector space [20]

Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a type of deep learning neural network that is generally used in the computer vision field but it is recently gaining popularity in the field of text processing due to it's ability to extract semantic meaning from feature sets and due to it's ability to process multi-dimensional feature vectors. CNN is a type of Artificial Neural Network(ANN), these neural networks are inspired by the Biological Neural Networks found in the animal brain, they consist of a network of neurons that are also called nodes, these nodes are fed an input and they perform mathematical operations on that input and pass on the output to the next neuron. A convolutional neural network is made up of three layers that communicate with each other, these are

- the convolutional layer, the pooling layer and the fully-connected layer. These three layers stacked together form a Convolutional Neural Network[10]. There have been some studies that use Artificial and Convoloutional Neural Networks to identify fake reviews. Zhao et all.[21] used a word order preserving optimised CNN to detect fake reviews and achieved 70% accuracy.

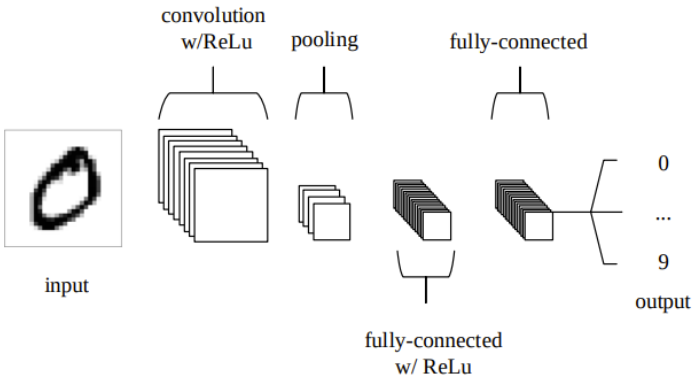

FIGURE 2.7: Architecture of a CNN Model [10]

Random Forest Classifiershttps://www.overleaf.com/project/6342e215f81be07375141b07 Random Forest Classifiers is a type of supervised learning classifier that uses a random subset of features, trains them on decision trees and then averages the predictions generated by these trees to create a final prediction. Random forests are used to reduce a models variance and prevent overfitting by using bootstrap sampling i.e. Using random rows from the data to train the model so that the model does can ignore the noise generated while training. Banerjee et al.[22] and Zhang et al.[23] have previously used Random Forest Classifiers to detect fake reviews.

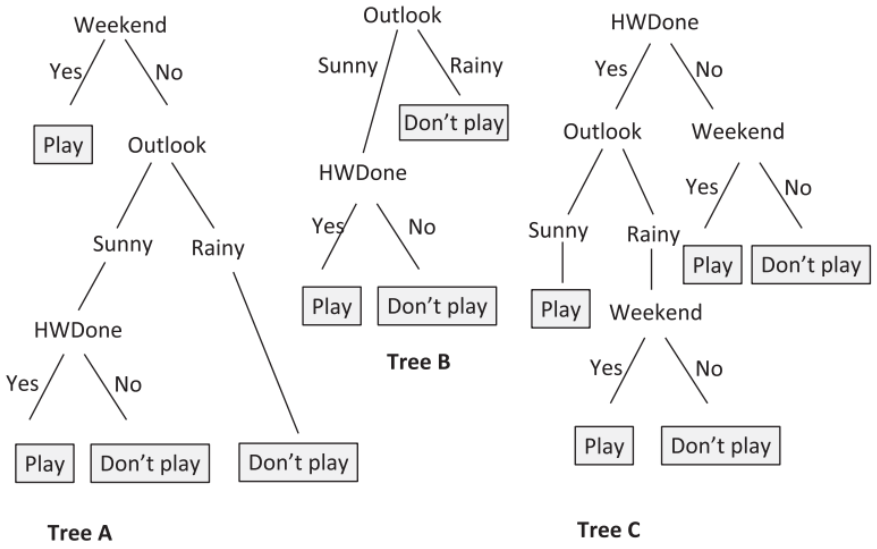

FIGURE 2.8: Random Forest of three decision trees [24]

2.2.3.2 Topic Modelling

Topic modelling is an unsupervised Natural Language Processing technique. A topic model consists of a family of probability distributions over a set of topics extracted from a collection of text, i.e. a topic modelling algorithm mines clusters related to a specific topic based on word frequency where a topic is defined as a set of words with similar overall meaning[12]. There are three topic modelling algorithms -

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a topic modelling algorithm that does not use any Natural Language Processing nor does it use any resources created by humans like dictionaries or thesauri. LSA uses a bag of words model which is used to map the words in a corpus to a matrix where rows represent terms and columns represent documents. It then uses this term-document matrix to try and uncover some similarity structures that are useful to solve the problem at hand[25]. The LSA analysis consists of four mane steps, these are -

1. Term-Document Matrix - A bag of words model is used to map the words in a corpus to a matrix where rows represent terms and columns represent documents and the frequency of a term in a document is stored in the individual cells.

2. Transformed Term-Document Matrix - Entries in the term-document matrix are transformed. In this step, frequencies are added in a sub linear fashion and inversely with the overall occurrence of the term in a collection[25].

3. Dimension Reduction - This is the key idea of topic modelling - to map documents to a vector space of reduced dimensionality. In Latent Semantic Analysis, **Singular Value Decomposition(SVD)** is performed on the matrix in which the k largest singular values from the transformed term-document matrix are retained where the resulting matrix is the best k-dimensional approximation of the original matrix [25].

4. Retrieval in Reduced Space - Similarities are computed among the entities in the reduced dimensional vector space.

Probabilistic Sentiment Analysis

Probabilistic Latent Semantic Analysis (pLSA) is a topic modelling algorithm that substitutes the Singular Value Decomposition technique mentioned above with the Probabilistic model (also known as the aspect model shown) in equation 2.1. [26]

$$P(d,w) = P(d)P(w|d), P(w|d) = \sum P(w|z)P(z|d) \tag{2.1}$$

Latent Dirichlet Allocation

Latent Dirichlet Allocation(LDA) is a topic modelling algorithm that is the bayesian version of pLSA. It's basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The probability density of a k-dimensional Dirichlet random variable that can take values in the (k - 1) simplex (a k-vector lies in the (k1)-simplex if i 0, k i=1 i = 1),has the following probability density in this simplex[27].

$$P(\theta \mid \alpha) = \frac{\gamma(\sum_i^k \alpha_i)}{\pi_i^k \gamma(\alpha_i)} \left(\theta_i^{\alpha-1)} ... \theta_k^{\alpha-1)}\right) \tag{2.2}$$

2.2.3.3 Overview

Detecting fake reviews using topic modelling is a study that has not been explored too much yet. Birim et al.[12] used topic modelling to detect fake reviews and achieved an accuracy of 81% which is one of the only studies that could be found on detecting fake reviews using topic modelling. For this reason, this study will use topic modelling to verify the findings of [12] and to reason whether the field of spam detectio using topic modelling is worth exploring or not. Also, as seen in table 2.1 the SVM and the Naive Bayes algorithms have been used very often when it comes to doing this study and they consistently achieve respectable results. Therefore, the Naive Bayes algorithm and the SVM algorithm will be used to implement the research conducted in this paper. Finally,

Word2Vec is one of the most popular programs for word embedding, it is also available in Scikit-learn which is the scientific library being used to conduct the research in this paper. Therefore, Word2Vec will also be used for word embedding and semantic parsing.

Chapter 3

Problem - Malicious review detection on online E-Commerce platforms

This chapter will discuss the problem definition and the intended method to achieve the solution to the problem laid out. In this chapter, we will also lay out the functional and non-functional requirements to implement the methods discussed.

3.1 Problem Definition

Opinion reviews are one of the most important factors in a customers decision making process when it comes to purchasing products online. A lot of bad actors use this to their advantage by review bombing their victims. A review bomb is an Internet phenomenon consisting of a massive attack by groups of Internet users on a website that displays users' reviews[28]. Businesses review bomb their competition to gain an edge , consumers review bomb products that they agree with, businesses buy fake reviews to inflate the value of their own products etc. There have been several high profile reports of video games and movies and other products being review bombed and the reports show no signs of letting up. Due to malpractices like these still being practiced, preserving the integrity of reviews is a very important problem to solve to have an online marketplace that is fair for all and productive by encouraging competition and incentivising making good products for the consumer.

3.2 Objectives

Several researches have looked to solve this problem using supervised learning methods and have achieved some success but these models are all trained on different datasets which makes it harder to determine the techniques that are most efficient. The main objective of this paper is to compare the two most popular datasets for e-commerce reviews i.e. The Amazon dataset and the Yelp dataset and to posit the best dataset out of the two for identifying opinion spam on e-commerce websites. this paper will also expand on previous work done by building and training a supervised learning model using techniques and algorithms that have proved most successful in identifying fake reviews previously.

3.3 Functional Requirements

In this section, we will discuss the functional requirements to build and test a classifier that can detect fake reviews on e-commerce platforms.

Dataset

Having a good dataset is one of the most important parts of creating an accurate supervised label warning. A dataset is considered a good quality dataset for training a machine learning mode if -

1. It is accurately labeled - The dataset must have accurately labeled data as the supervised learning model will use these labels as their target variable and will be trained using these labels as the singular source of truth. These models will also use the labels in the dataset as the singular source of truth when checking the accuracy of the model

2. It has a lot of rows - A training dataset should have a lot of training instances as generally the more labelled rows in the dataset the better it is.

3. It has Variability - The more variety a dataset has, the more value that dataset can provide to the algorithm. Failure to learn train the model using data with more variety could lead to overfitting and poor performance when the model is faced with new scenarios. For example, a model that is trained based on daytime lighting conditions may show good performance on images captured in the day but will struggle under nighttime conditions whereas a dataset that has variety and has images for all times of the day will perform better for nightime conditions.Also, a model could be biased if a class is overrepresented in the dataset, for example, If

a review dataset has a lot more true reviews than fake reviews, there is a chance that the model trained using this dataset will be biased towards predicting true reviews.

Most of the peer-reviewed research papers that have been written before this paper have either used the **Amazon Review Dataset**[3][5][6][7][9] or the **Yelp Review Dataset**[4][8][9] for this reason, this paper will operate under the assumption that these datasets are the two best quality datasets available to perform this task and compare the results obtained by these two datasets to find which website has data that is better to train an algorithm that predicts real and fake reviews.

Programming Language and Scienific Library Although languages like Java and JavaScript have been used to build machine learning models, they are not as well documented, easy to use or efficient as python. Over the years python has established itself as the most popular programming language for machine learning thanks to it's ecosystem of scientific libraries. One of these libraries, Scikit-learn provides state-of-the-art implementations of many well known machine learning algorithms while maintaining an easy to use interface. Scikit-learn differs from other machine learning toolboxes in Python for various reasons -
i) it is distributed under the BSD license which is a free software license, which imposes minimal restrictions on the use and distribution of covered software, In other words, It is an open-sourced software.
ii) It incorporates compiled code for efficiency.
iii) While the package is mostly written in Python, it incorporates the C++ libraries LibSVM and LibLinear that provide reference implementations of SVMs and generalized linear models with compatible licenses[29].

3.4 Non-Functional Requirements

Requirements Engineering (RE) research suggests that considering Non Functional Requirements (NFR's) is critical to the success of a system. A non functional requirement is defined as any attribute or quality that is non functional but this broad definition is not always suitable[30]. To facilitate NFR's in machine learning, it is more useful to think about the usability of a machine learning model as the main entity rather than the model itself. Li et al. proposed a modeling language for non-functional requirements (NFRs) that views NFRs as requirements over qualities, mapping a software-related domain to a quality space[31]. Using this modelling language, we can come up with some Non Functional Requirements to build a Machine Learning algorithm. These Requirements are outlined below[32] -

1. **Accuracy and Performance** - One of the most important factors in evaluated a machine learning model is the models accuracy. The greater the accuracy with which a model can predict data it has not encountered before, the better the model is said to be. Algorithm performance, i.e. The time and space complexity that a model takes up is also an important factor in evaluating the quality of a model[32].

2. **Fairness** - The field of fair machine learning aims to ensure that decisions made by machine learning algorithms are equitable. Over the last several years, The main formal methods of implementing fairness in machine learning are - (1) anti-classification, meaning that protected attributes—like race, gender, and their proxies—are not explicitly used to make decisions; (2) classification parity, meaning that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes; and (3) calibration, meaning that conditional on risk estimates, outcomes are independent of protected attributes[33]. [32]

3. **Transparency** - In recent times, decisions made by machine learning models have gained great significance when it comes to making decisions affecting real human beings. But it is often not clear how these decisions were made. Machine learning models are not always objective and are prone to bias since the data injected in them may be inherently biased[34]. For this reason, to make sure that the algorithm is fair, the dataset and the process of creating and training the model should be transparent so that people can trust decisions made by the models. [32]

4. **Testing** - It is difficult to detect faults with machine learning algorithms because conventional software testing techniques are not always applicable in machine learning. It may particularly be difficult to detect subtle errors, defects or anomalies in many machine learning models because the only way to do that would be train a model on a completely different dataset which may be either impossible or too expensive to do [35].

Chapter 4

Implementation Approach

This section will outline the approach used to achieve the aims of this paper(outlined in 2). The architecture and methodology used to build the software outlined in this paper will be found in this section. **Note - Building and deploying a web application is also part of the scope of this project, no prior mention was made to this as it was not part of the research phase of the project. As the web application is part of the implementation, this section will involve steps to build and deploy the web application and link it to the machine learning algorithm.**

4.1 Architecture

To better explain the architecture of the software as a whole, it would be useful to divide the software into three entities according to how they would be deployed -

- Frontend
- Backend
- Machine Learning Algorithm API.

4.1.1 Frontend

The frontend of the Web Application will be implemented using HTML, CSS, VueJS and and Tailwind. AJAX will be used to communicate with the backend and the communication will take place using the HTTP 1.1 protocol. Vue JS uses component based architecture where components can be written in HTML, CSS, and JavaScript without

dividing them into separate files. Vue JS also uses a virtual DOM which provides applications built using it a huge boost in performance and scalability. A virtual DOM is a DOM that is created using JavaScript objects which imitates the real DOM of the webpage. When the virtual DOM changes, Vue compares the new and the old state and decides if the DOM needs to be updated. This process is called reconciliation. If a change is required only the associated DOM nodes will be altered while the rest of the tree remains intact. This allows Vue JS applications the ability to show changes in real time without having to refresh the page. AJAX is one of the most popular libraries used to make requests to a web server and tailwind grants the ability to write clean and scalable CSS webpages which is why it is being used.

4.1.2 Backend

The backend of the Web Application will be built using Maven, XML, Spring Boot and Hibernate. Spring Boot is an open-source platform built for Java developers to create Spring Applications easily. The advantages of using spring boot are -

- Spring boot provides "Inversion of control" i.e it takes over the control of objects or portions of a program. This helps developers to reduce errors and to reduce development time.
- Spring boot provides a powerful batch processing and manages REST endpoints.
- It provides a flexible way to configure Java Beans, XML configurations, and Database Transactions.
- In Spring Boot, everything is auto configured; no manual configurations are needed.
- offers annotation-based spring application
- Eases dependency management

the database that will be used will be PostgreSQL. SQL allows for faster and more efficient query processing as it is as it stored data in a structured way which allows for easier fetching and altering of data. In relational databases, CRUD operations like insertion, deletion, querying, manipulation can be accomplished in a matter of seconds. The backend will be deployed to the internet using docker and a cloud platform such as Heroku or Google Cloud Platform.

4.1.3 Machine Learning Algorithm

The machine learning algorithm will be built using scikit-learn, numpy and pandas. The algorithms that will be used have been outlined in the overview section in 2. Flask will be used to deploy the trained algorithm as an API so that it can communicate with the backend of the web-application.

4.2 Risk Assessment

TABLE 4.1: Initial risk matrix

Frequency/ Consequence	**1-Rare**	**2-Remote**	**3-Occasional**	**4-Probable**	**5-Frequent**
4-Fatal	1				
3-Critical		2			
2-Major	3				4
1-Minor					

1. **Unprocurable Dataset**
    - Fatality - Fatal
    - Chance of occurring - Rare
    - Explanation - Every classification algorithm needs a labelled dataset for it to be trained on. If we would not have found any real world labelled datasets of fake reviews it would not be possible to identify fake reviews using machine learning.
    - Mitigation - Obtained two real world datasets, one of which has been used in similar research before which makes it easier to evaluate findings obtained when using that dataset as well.
2. **Hardware Issues**
    - Fatality - Critical
    - Chance of occurring - Remote
    - Explanation - Training a machine learning algorithm is an expensive task, especially when using unsupervised algorithms. It would be possible that a machine runs out of RAM when trying to train an algoriithm.
    - Mitigation - there is only a remote chance of the issue occuring is sci-kit learn is a scientific library that contains stae of the art algorithms that are very

efficient. In case of it occurring, a cloud platform can be used to train the machine learning algorithm and perform the research.

3. **Loss of data**
    - Fatality - Major
    - Chance of occuring - Rare
    - Explanation - When in development, there is a chance that the work done on the project could get lost.
    - Mitigation - Will be using Github for the purpose of source control and will update the remote repository regularly.
4. **No deployement platform**
    - Fatality - Major
    - Chance of occuring - Frequent
    - Deploying a microservice requires a platform that will host the code on the internet. It could be diffiult to find a platform that does this for free.
    - Mitigation - Heroku is a free platform to deploy code, in case of that being unavailable, we will pay to host or microservice for a certain period of time.

4.3 Methodology

This section will outline how the research described in the paper will be implemented and evaluated, the first section will outline a more general approach to solving the problem and then describe my personal approach and the skills I will need to make the implementation possible.

4.3.1 Required Steps

This section will describe the steps required to implement a machine learning algorithm that can predict fake reviews.

1. **Data Collection** - For the purpose of this paper, two datasets have been collected, the yelp dataset is labelled and was used in the paper [4] by Barbado et al. This dataset contains reviews found on Yelp.com and are labelled real or deceptive. The second is a unlabelled dataset containing reviews found on Amazon.com, these reviews are not labelled and the one of the objectives of the paper is to try and

attempt to use an unsupervised learning technique called topic modelling to try and label this dataset.

2. **Data Cleaning** - In machine learning, data cleaning is the process of removing redundancies, errors and modyfing data that is incorrect to prepare a dataset to be trained by a machine learning model. Data Cleaning is used to remove instances from a dataset that would be removed in the real-world like blankspaces, spelling errors etc.

3. **Operationalization of sentimental clues(Classification)** - As indicated earlier, this paper attempts to distinguish authentic and fake reviews based on sentimental clues, mainly based on how many positive and negative words exist in real and in fake reviews. This paper will also the psycholinguistically motivated features and bigram laid out by [7].

4. **Dimensionality Reduction (Clustering)** - The Amazon dataset is not labelled, therefore labelling the reviews in the dataset as true or deceptive will have to be done by an unsupervised learning model. The Latent Dirichlet Allocation(LDA) algorithm that was also used by [12] will be used to attempt to label the dataset.

5. **Evaluation** - The final step will be to evaluate the algorithm that is developed by comparing previous results and investigating Non Functional Requirements(See Evaluation).

4.4 Evaluation

As mentioned in the Section 3.4 in chapter 3, research in requirements engineering has concluded that when it comes to machine learning, it is more useful to think about the usability of a machine learning model as the main entity rather than the model itself. Therefore, according to the framework proposed by Li et al.[31], the evaluation of the Non-Functional Requirements of the model would be the best way to evaluate the quality of the model itself. The Non-Functional Requirements that will be evaluated in this paper are -

1. **Accuracy and Performance** - The evaluation of the accuracy and performance of the model developed will be done using a confusion matrix and k-fold cross validation. k-fold cross validation is a standard procedure in machine learning used to evaluate the accuracy score of a model. K-fold divides the data set into n number of train and test folds, train and test folds are then picked randomly from the data set. This is done so that the data is trained and tested on all the

data and to prevent over fitting. The number of true positives, true negatives, false positives, and false negatives obtained from each iteration of the k-fold split are then stored in a matrix called the confusion matrix. Evaluating the confusion matrix gives us the accuracy score of a machine learning model.

2. **Transparency** - The dataset used in this research are available online and the methods used in this paper will be properly documented so that any researcher can verify the findings postulated by this paper ensuring full transparency.

When it comes to evaluating the website, the main factors will be -

1. Functionality - Whether the website works or not
2. Features - The features implemented in the website
3. Design - The "look and feel" of the website
4. Performance - How fast the website can load it's components.

4.5 Prototype

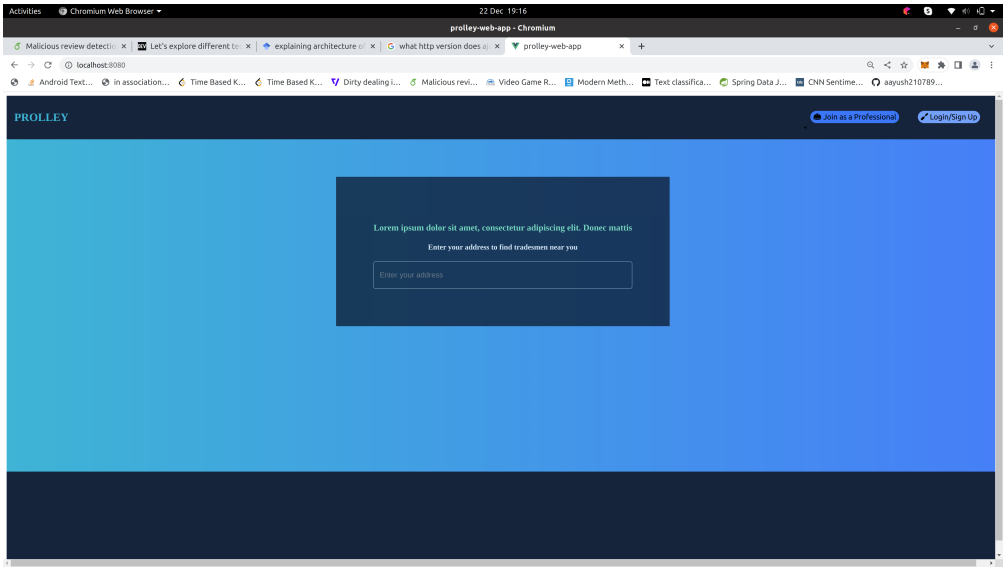

FIGURE 4.1: Prototype of the Frontend for the website

Chapter 5

Implementation

This chapter presents the implementation details of the approach proposed in 4 for detecting opinion spam in e-commerce reviews. It provides insights into the actual implementation process, highlighting the challenges encountered along the way and a comparison between the actual solution approach and the initial proposed approach.

5.1 Difficulties Encountered

The different difficulties that were found when developing the solution approach are listed below, these are grouped into three categories easy, medium and hard. The easy problems were solved with little difficulty while the medium and hard problems were more challenging and required making changes to the initial implementation approach.

5.1.1 Difficulty - Hard

1. Description: Finding out whether topic modeling can be used to group reviews into fake and true categories was challenging due to limited documentation and previous research in this area.

   Impact on Project Design:

   - Architecture: The original solution approach relied on existing topic modeling techniques or methodologies, which were not readily available.
   - Project Risk: The difficulty represented a risk as it affected one of the problem statements of the project.
   - Methodology: The challenge required devising a unique solution that combined topic modelling and traditional supervised learning models instead of

leveraging established approaches but an algorithm was ultimately created, therefore the methodology of the project was not changed.

- Implementation Schedule: The difficulty caused a lot of delays in the implementation schedule as a lot of trial and error was involved in trying to find out whether clustering can be used to label an unlabelled dataset
- Evaluation Plan: The original evaluation plan assumed the availability of established topic modeling techniques. These could not be found, therefore the evaluation plan had to be changed, the new evaluation plan will be presented in the next subsection.
- Management of Difficulty: To overcome this difficulty, a unique solution was developed, which will be presented in the paper. This involved designing and implementing topic modelling and add it's results as a feature vector for a supervised learning model

2. Description: The initial implementation approach involved creating a website to send data to the machine learning models. However, the time spent on researching the topic modeling left insufficient time to create the website.

   Impact on Project Design:

   - Architecture: The original project design assumed the inclusion of a website as a component for data input and interaction with the machine learning models.
   - Project Risk: The difficulty represented a significant risk as it compromised the planned functionality and usability of the solution.
   - Methodology: The inability to create the website impacted the methodology, as the project had to adapt to alternative data input methods such as csv and json files.
   - Implementation Schedule: The lack of time for website development freed up time in the implementation schedule to develop a topic modelling algorithm that works.
   - Evaluation Plan: The original evaluation plan had considered the website's interaction and user experience as part of the evaluation criteria, but this had to be adjusted.
   - Management of Difficulty: To manage this difficulty, the focus shifted to alternative data input methods, such as using pre-collected datasets allowing for the evaluation of the machine learning models' effectiveness without the website component.

5.1.2 Difficulty - Medium

1. Description: The lack of documentation on using only text data to analyze reviews made it challenging to determine the viability of this approach.

   Impact on Project Design:

   - Architecture: The original project design assumed the availability of research for only text-based review analysis but except for the work done by ott et al. [10] no other papers could be found that had done analysis **only** on the text of the review.
   - Project Risk: The difficulty represented a risk as the effectiveness of the proposed approach depended on the viability of text analysis alone.
   - Methodology: The methodology of the project did not need to be changed as there is a lot of documentation on supervised learning and sentiment analysis and the evaluation of these algorithms is well mapped out by now.
   - Implementation Schedule: The additional effort required to explore the viability of the algorithms that were written caused small delays in the implementation schedule.
   - Evaluation Plan: A new metric called **f1-score** was added to evaluate the effectiveness of the model.
   - Management of Difficulty: To address this difficulty,techniques such as confusion matrices and k-fold cross verification were used and the evaluation plan was adjusted to add a f1-score metric.

5.1.3 Difficulty - Easy

1. Description: The real-world yelp dataset obtained had a class imbalance, with a notably higher number of real reviews compared to fake ones

   Impact on Project Design:

   - Architecture: The original project design did not account for the class imbalance issue in the dataset.
   - Project Risk: The difficulty represented a risk as the effectiveness of the proposed approach depended on the viability of text analysis alone.
   - Methodology: The methodology of the project needed to be changed a little to include metrics such as precision recall and f1-score to get a better idea of the effectiveness of the model.

- Implementation Schedule: Learning about and researching class imbalances and what to do about them took some time off the implementation schedule
- Evaluation Plan: A new metric called **f1-score** was added to evaluate the effectiveness of the model. Along with precision and recall.
- Management of Difficulty: To address this difficulty,techniques such as confusion matrices and k-fold cross verification were used and the evaluation plan was adjusted to add a f1-score metric.

5.2 Actual vs Predicted Solution Approach

This section will compare the predicted solution approach in Chapter 4 with the actual solution approach. More information on the final methodology and the system testing can be found in Chapter 6

1. **Architecture**

   Difference: The original design included a frontend, backend, and machine learning algorithm API as separate entities in the architecture. However, the final project omitted the frontend and backend components.

   Justification: The web application, from the beginning was not a critical goal of this project as it was not very relevant to the research questions and the objectives laid out in chapter 3. The difficulty in creating topic modelling algorithm, as mentioned earlier, led to the the decision being made to focus solely on the machine learning algorithm implementation, which was deemed more critical for achieving the project's objectives within the given constraints. Furthermore, the initial goal to build a website and deploy it on the internet was made to get some more real world data to work with. But that data would be unlabelled and not very valuable, we found an amazon dataset which was also unlabelled, the ready availability of this data made it a more viable alternative to use.

2. **Use cases**

   Difference: The primary use case of the implementation of this project was to answer the research questions that had been laid out earlier. This use case was satisfied as both the supervised learning and the unsupervised learning algorithms provided satisfactory results that helped in answering these questions. These results will be laid out in chapter 6

3. **Risk Assessment**

Difference: The risk associated with not having a deployment platform for the web application was realized in the final project.

Justification: The difficulty of not finding a suitable deployment platform for the machine learning algorithm api contributed to the decision of omitting the website from the final project as looking for a deployment platform and learning how to deploy an api on it was not viable given the time constraints.

4. **Methodology**

   Difference: The methodology as listed in chapter 4 involved five steps which were - Data Collection, Data Cleaning, Operationalization of sentimential clues, dimensionality reduction and evaluation. This methodology was followed to obtain the results of this research and no changes had to be made as the intial solution approach was effective in achieving the research objectives.

5. **Evaluation Plan**

   Difference: The initial evaluation plan focused on checking the accuracy of the model using k-fold cross-validation and a confusion matrix. However, during the implementation phase, it was discovered that the dataset had class imbalances, with a significantly higher number of real reviews compared to fake reviews. This class imbalance posed a challenge in accurately assessing the model's performance solely based on accuracy.

   Justification: The introduction of the F1 score metric, precision, and recall was necessary to address the class imbalance issue and obtain a more comprehensive evaluation of the model's performance. The F1 score takes into account both precision and recall, which are particularly useful when dealing with imbalanced datasets. Precision measures the proportion of correctly identified fake reviews among all predicted fake reviews, while recall measures the proportion of correctly identified fake reviews among all actual fake reviews. By considering both precision and recall, the F1 score provides a more balanced assessment of the model's performance, especially in cases where the class distribution is imbalanced.

   Additionally, the evaluation plan aimed to ensure transparency by making the datasets used in the research available online and properly documenting all the work done. This approach allows other researchers to verify the findings and promotes transparency in the research process. By providing access to the datasets and documenting the methodology, researchers can replicate the experiments and validate the research outcomes, ensuring the credibility and transparency of the model and its implementation.

6. **Prototype of the resulting product**

Difference: The initial paper included a prototype of the frontend of the website, showcasing the design and functionality of the user interface. However, in the final project, the prototype of the frontend was not implemented due to time constraints.

Justification: The decision not to implement the frontend prototype in the final project was based on time constraints. Developing a fully functional web application requires significant time and resources, which were limited within the scope of the project. Prioritizing the machine learning algorithm was necessary to ensure the successful completion of the project within the given timeframe.

However, to provide a reference for future work and to demonstrate the intended design and functionality, the frontend prototype was pushed to GitHub. By sharing the prototype on GitHub, it remains accessible to interested parties who may wish to build upon the project or further develop the frontend in the future. This allows for potential continuation and improvement of the project beyond the scope of the current implementation.

1

Chapter 6

Testing and Evaluation

In this chapter, an objective evaluation of the final models will be conducted, focusing on quantitative analysis. The evaluation aims to provide a comprehensive assessment of the system's performance and effectiveness, comparing it, where possible, to commercially available fake review detection models and to models described in research papers.

To provide context to this chapter, a brief recap of the problem statement, research questions and and the requirements of this project will be needed. As mentioned in chapter 1 the main objectives of this paper were to -

1. Compare the Amazon dataset and the Yelp dataset for e-commerce reviews.
2. Determine the best dataset for identifying opinion spam on e-commerce websites.
3. Build and train a supervised learning model using techniques and algorithms that have been successful in identifying fake reviews.

To achieve the objectives of the paper, several different machine learning models were created, and an objective evaluation was conducted. As mentioned in chapter 4 the metrics for this evaluation were -

6.1 Metrics

Functional Requirements

Functional requirements are those requirements that outline specific criteria and constraints that the model evaluation must satisfy in order to meet its intended purpose effectively. The functional requirements used to evaluate the results of this project are -

1. Data Quality and Quantity Requirements -

   - Accuracy of the dataset: The yelp dataset contained accurately labelled real world data but the tradeoff was that it only had 30,000 instances.
   - Sufficient number of rows: The yelp dataset had 30,000 instances to train the model on while the Amazon dataset contained 21,000 data points. Considerations to this were made when comparing the two datasets to each other.
   - Variability in the dataset: The text of the reviews were the only features used in this research, and because these were real world reviews, there was a good amount of semantic variability in the corpus.

2. Consideration of class imbalance - The yelp dataset had 20828 reviews labelled real and only 9653 reviews labelled false. Accuracy is not a good metric when there is a class imbalance in the dataset therefore f1 score was intorduced to get a better measure of the effectiveness of the model. Furthermore, the Naive Bayes and AdaBoost algorithms were used which are both robust to overfitting especially when trained on low noise data[36].

3. Supervised Model Evaluation Metrics -

   - F1 Score - Accuracy is a widely used metric for measuring the performance of a classifier, however, when the prior probabilities of the classes are very different, this metric can be misleading. A better choice is F1-score, which can be interpreted as a weighted average of the precision and recall values[37]: $F_1 = 2.\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
   - Accuracy - Although accuracy is a slightly inaccurate metric when it comes to skewed datasets, it provides a good baseline for a models performance therfore it is useful to measure it. Accuracy is the the ratio of correct predictions made by a model and all predctions made by the model. $Accuracy = \frac{\text{TP}+\text{TN}}{\text{TN}+\text{TP}+FN+FP}$
   - ROC Curve - The ROC curve depicts the true positive rate as the function of the false positive rate

   Jeni et al. in their research found that F1 Score and ROC curve are the metrics that show the least skew on examining binary SVM models trained on expert annotated datasets[37].

4. Topic Modelling Evaluation Metrics -

   Evaluating the effectiveness of a topic modeling algorithm involves assessing the quality of the generated topics and their coherence with the underlying data. Some of the approaches that were considered are -

- Perplexity - Perplexity is a widely used metric to evaluate topic models, particularly those based on probabilistic modeling like Latent Dirichlet Allocation (LDA) the perplexity is calculated by first estimating the topic proportions and word distributions from the topic model, and then computing the probability of each word in the test set given these estimated distributions. The formula sums up the log probabilities of all the words in the test set and exponentiates the negative sum to get the perplexity score - $Perplexity = exp(-\sigma(p(w)log(q(w))))$ Hoowever, according to some research [38] Perplexity is unclear when used to evaluate the quality of topic modelling from the perspective of human judgement. therefore perplexity was not used in order to evaluate the topic modelling algorithm.
- Normalized Pointwise Mutual Information(NPMI) - According to studies in the field of topic modelling, NPMI has been found to the highest correlations with human interpretation scores [12]. Therefore in this paper, NPMI was used as the evaluation metric. NPMI is calculated using the formula - $NPMI(w_i, w_j) = \frac{PMI(w_i,w_j)}{-logp(w_i,w_j)}, i \neq j$ where $PMI(w_i, w_j) = log(\frac{p(w_i,w_j)}{p(w_i)p(w_j)}), i \neq j$ [12]

6.2 System Testing

The implementation of the project was divided into three phases, the first phase involved performing natural language processing on the yelp dataset and training supervised learning models to classify a review as real or fake. The second more ambitious phase involved checking whether it is possible to label an unlabelled dataset using a clustering technique called topic modelling. As mentioned in chapter 4 several attempts were made to create a topic modelling algorithm that can cluster reviews into a real reviews cluster and a fake reviews cluster but these attempts were unsuccessful which led to the conclusion that labeling an unlabelled dataset using unsupervised training methods alone is extremely challenging, if not impossible. This led to phase three where a new approach was tried,incorporating the results of topic modelling into a supervised model for labeling the unlabelled dataset.

6.2.1 Phase1 - NLP on yelp dataset

In this phase, a real-life Yelp dataset containing labeled real and fake reviews was used to perform supervised learning using Naive Bayes, AdaBoost, and SVM algorithms. The primary feature utilized in the analysis was the content of the reviews. The results

of the experiments showed that Naive Bayes achieved an accuracy of 67% and an F1 score of 78%. The SVM algorithm achieved an accuracy of 62.70% and an F1 score of 78.7%. AdaBoost achieved an accuracy of 67.8% and an F1 score of 80.3%. It is worth noting that the F1 score was significantly higher than the accuracy for all three algorithms. The discrepancy between the F1 score and accuracy is attributed to the class imbalance present in the dataset. The dataset consisted of 20,828 real reviews and 9,653 fake reviews. In situations with class imbalance, accuracy can be misleading, while the f1 score provides a better representation of a classifier's performance. Class imbalance occurs when one class has a significantly larger number of instances than the other. In this case, the presence of more real reviews than fake reviews leads to a higher F1 score. The F1 score combines precision and recall, which makes it less sensitive to class imbalances compared to accuracy. Precision is the ability of the model to correctly identify positive instances, while recall measures the model's ability to find all positive instances. These accuracy results are consistent with previous research by Ott et al., who also found a maximum accuracy of 67% when training models on a real-life Yelp dataset.

In addition, natural language processing techniques were used to gain insights from the data. The analysis showed that fake reviews had a much higher probability of having prepositions, while real reviews had a much higher probability of using conjunctions. This difference could be because fake reviews are often written to manipulate or deceive readers and may use simpler language structures to avoid detection. On the other hand, real reviews may use more complex language structures because they are written by genuine customers who want to provide detailed and informative feedback.

Another observation made was that real reviews had a significantly higher word count than fake reviews. This observation is consistent with previous research on the topic, which suggests that real reviews are more detailed and provide more information than fake reviews. The reason for this could be that real reviews are written by genuine customers who have had real experiences with the product or service and want to provide a comprehensive review, while fake reviews may be written by bots or paid reviewers who do not have personal experience with the product or service.

Overall, the research done in this phase showed that machine learning algorithms can be used to distinguish between real and fake reviews with moderate accuracy when trained only on the content of the review as a feature. Natural language processing techniques can also provide valuable insights into the language patterns of real and fake reviews.

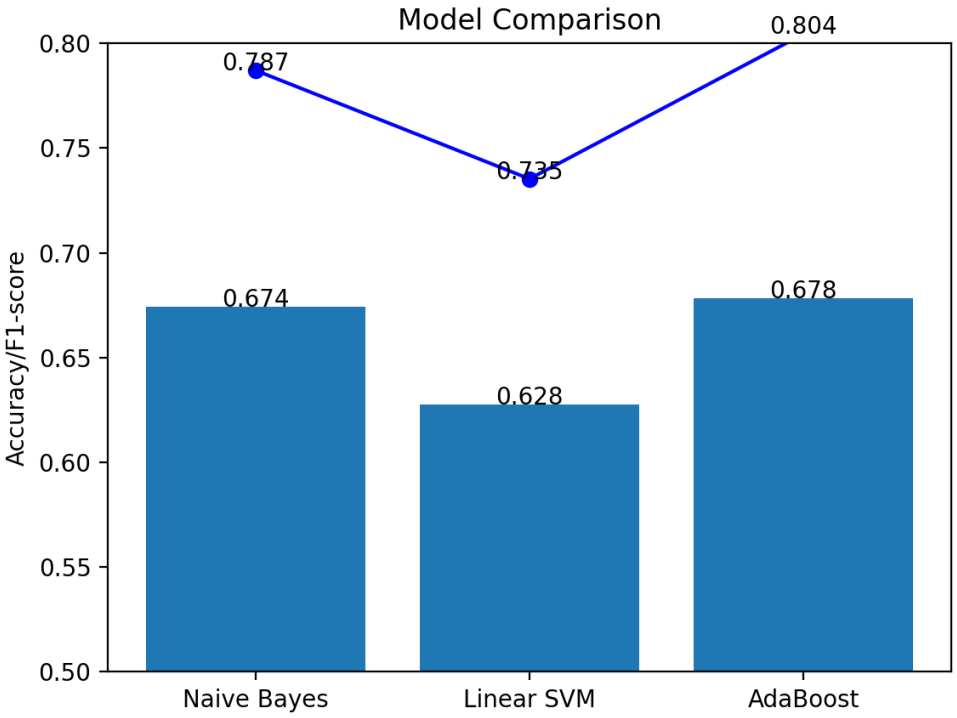

FIGURE 6.1: Comparision on accuracies and f1 scores when only review text was used as a feature on the yelp datatset.

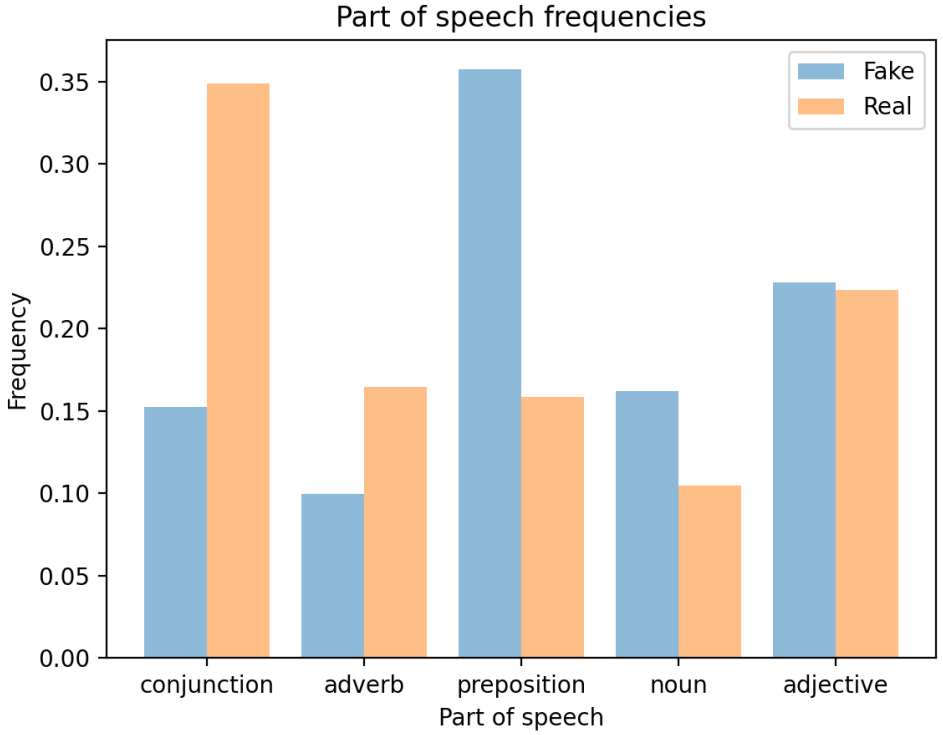

FIGURE 6.2: Comparision of part of speech frequencies in real and fake reviews.

6.2.2 Phase2 - Attempt to cluster fake reviews using topic modelling

In this research paper, the initial approach was to use topic modeling, specifically Latent Dirichlet Allocation (LDA), to cluster all the fake reviews and real reviews together. The assumption was that LDA, being a popular technique in topic modeling, would be able to differentiate between real and fake reviews based on the topics discussed in the reviews. However, upon further exploration and understanding of topic modeling, it was realized that this assumption was flawed.

Topic modeling, such as LDA, aims to identify topics by grouping words that are more likely to occur together. It does not inherently distinguish between real and fake reviews. Fake reviews can utilize any combination of words, whether likely or unlikely to co-occur, to manipulate or deceive readers. Therefore, the assumption that topic modeling alone would cluster real and fake reviews was based on a flawed understanding of the capabilities of topic modeling.

Consequently, the initial aim of the research, which was to use natural language processing and natural language understanding to create an unsupervised learning model using topic modeling and label the Amazon review dataset as "Genuine" or "Fake," had to be altered. Instead, a semi-supervised approach was adopted to label the Amazon dataset.

The revised approach involved finding a dataset of Amazon reviews that belonged to a single topic. Topic modeling was then used as a feature, along with other useful features discovered in phase 1, such as review length and part-of-speech frequencies, to train a supervised learning model. The most effective model identified in phase 1, AdaBoost, was utilized for this purpose.

By combining topic modeling as a feature with other informative features, the new approach aimed to detect spam or fake reviews. Computer generated reviews that were off-topic would stand out as potential fakes. This hybrid approach allowed for a more accurate labeling of the Amazon review dataset by leveraging both topic modeling and supervised learning techniques.

Overall, this adjustment in the research methodology demonstrates the importance of adapting the approach based on a deeper understanding of the underlying techniques. By incorporating insights gained from phase 1 and adopting a semi-supervised approach, the research aimed to improve the accuracy of distinguishing genuine and fake reviews using a combination of natural language processing, topic modeling, and supervised learning methods.

6.2.3 Phase3 - Use a semi-supervised approach to label real and fake amazon reviews

In the subsequent phase of the research, a labeled Amazon dataset was obtained, which encompassed reviews of various types. Several features were extracted from this dataset to train an AdaBoost classifier. These features included the review length, the most likely topic associated with the review, and the sentiment score of the review. The classifier achieved an accuracy of 66% when evaluated using a train-test split, where the data used for testing was a part of the same dataset.

It was observed that since the dataset contained reviews of diverse topics, the incorporation of topic modeling did not significantly impact the accuracy of the classifier. Topic modeling aims to identify underlying topics within a corpus by grouping words that are more likely to co-occur. However, in the context of the mixed-topic dataset, the variability of topics within each class (real or fake) diminished the distinguishing power of topic modeling. As a result, the contribution of topic modeling to the overall accuracy was limited.

Furthermore, an additional feature was introduced, namely, whether the reviewer was a verified purchaser or not. Although this feature was initially beyond the scope of the review-centric features, it was included in the study for experimental purposes. Surprisingly, the addition of this reviewer-centric feature led to a significant increase in accuracy, reaching 80%.

The improved accuracy achieved by incorporating the reviewer-centric feature can be attributed to the fact that verified purchaser status provides credibility to the review. Previous research conducted by Mukherjee et al. and Barbado et al. supports this finding, highlighting the efficacy of reviewer-centric features in detecting spam. Reviews from verified purchasers are more likely to be authentic and reliable, making them valuable indicators for differentiating between genuine and fake reviews.

Overall, this phase of the research demonstrated that while topic modeling did not greatly enhance accuracy in a mixed-topic dataset, the inclusion of reviewer-centric features, such as the verified purchaser status, played a crucial role in improving the classifier's performance. The findings align with previous studies, underscoring the significance of reviewer-centric features in spam detection tasks.

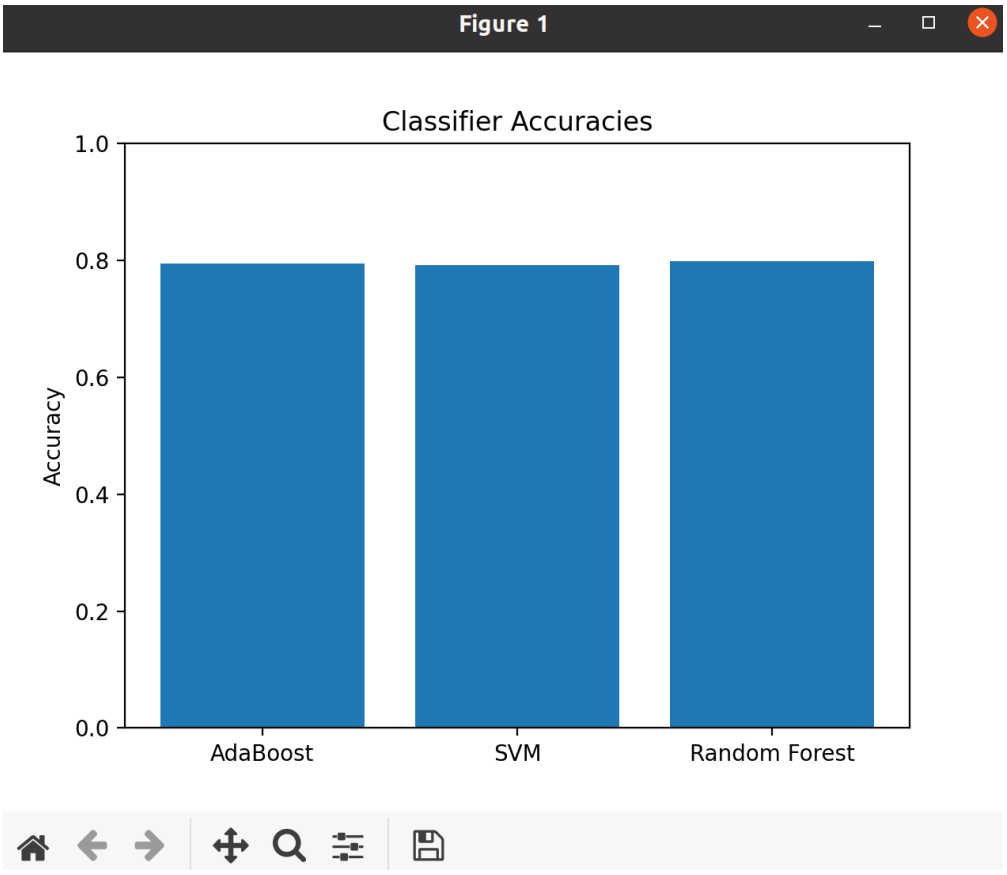

FIGURE 6.3: Accuracies of top three models with topic, review text and verified purchase as features

6.3 Results

In this study, extensive research was conducted using various machine learning models trained with different features. The results of each phase are presented below.

6.3.1 Yelp Dataset

Phase 1 involved training several supervised learning models on the yelp data set to find the best models and features for phase 2. When evaluating the models, it was found that they produced significantly higher recall scores as compared to precision scores.

The significantly higher recall scores in table 6.1 can be attributed to the class imbalance in the dataset. When a dataset has a class imbalance, models trained on it tend to predict the majority class more frequently than the minority class. In this case, since there are many more real reviews than fake reviews, the models tend to predict real reviews more

frequently, resulting in a higher recall for the minority class (fake reviews) but a lower precision.

Recall measures the proportion of actual positives that are correctly identified by the model, whereas precision measures the proportion of predicted positives that are actually positive. In this case, the models are correctly identifying a high proportion of fake reviews (high recall) but also predicting many real reviews as fake (low precision). This is a common issue in class imbalance and can be addressed by techniques such as oversampling the minority class or using weighted loss functions during training to give more weight to the minority class.

Model	Features	Accuracy	Precision	Recall	F1 Score
Naive Bayes	RT	67.40	70.93	88.38%	78.70%
	RT + TD	67.75	71.22	88.31%	78.85%
	RT + TD + SA	-	-	-	-
SVM	RT	62.73%	71.24%	75.90%	73.50%
	RT + TD	67.57	**71.25%**	87.78%	78.66%
	RT + TD + SA	68.08%	68.08%	99%	81.01%
AdaBoost	RT	67.83%	68.63%	97.15%	80.44%
	RT + TD	68.20%	68.88%	97.20%	80.63%
	RT + TD + SA	68.67%	69.47%	96.36%	80.74%
Random Forest	RT	68.96%	69.73%	96.17%	80.84%
	RT + TD	69.33%	69.36%	98.43%	81.38%
	RT + TD + SA	**69.51%**	69.32%	**99.03%**	**81.56%**

TABLE 6.1: Supervised Learning Results On yelp dataset

it was observed that the accuracy of the best model, Random Forest, using only review text as a feature was 68%, which aligned with the findings of Ott et al. who achieved 67% accuracy on a real-world dataset. In an attempt to increase the accuracy of the model using solely NLP-based features, three NLP features: sentiment score, topic distribution, and length of the review, were added. However, as depicted in table 6.1, these features had little to no impact on the RF, NB, and AdaBoost models, whose accuracies only increased slightly. On the other hand, adding topic distribution increased the accuracy of the SVM model significantly.

It may be the case that topic modelling could have increased the accuracy, as the dataset contained reviews related only to hotels. Thus, the prior assumption that adding topic distribution as a feature to reviews that all relate to one topic could increase accuracy was proven correct, at least for this dataset and only for the SVM model.

One possible explanation for why adding topic distribution as a feature increased the accuracy of only the SVM model significantly, but not the other models, is that the SVM model is better at handling high-dimensional data. Topic distribution adds another

dimension to the feature space, which may have made it easier for the SVM model to differentiate between fake and genuine reviews.

Another possibility is that the SVM model is more sensitive to the specific structure of the data and the added feature improved its ability to capture that structure. In contrast, the other models, such as Random Forest and Naive Bayes, may have been less sensitive to the added feature or may have been better at handling the data with fewer dimensions.

6.3.2 Amazon Dataset

In the second phase of the research, the top three models from phase 1 were trained using an Amazon dataset containing 10,500 real reviews and 10,500 fake reviews, resulting in a balanced dataset with no class imbalances. Therefore, accuracy was an appropriate metric for evaluating the models' performance. As observed in the previous Yelp dataset, introducing topic density as a feature improved the SVM model's accuracy, while the random forest and AdaBoost models showed minimal or no improvement. The maximum accuracy achieved using only review-centric features was 67.69% by the SVM model.

Model	Features	Accuracy
SVM	RT	62.07%
	RT + TD	67.69%
	RT + TD + SA	67.54%
	RT + TD + VP	79.19%
AdaBoost	RT	63.54%
	RT + TD	63.69%
	RT + TD + SA	63.85%
	RT + TD + VP	79.40%
Random Forest	RT	66.11%
	RT + TD	67.23%
	RT + TD + SA	66.09%
	RT + TD + VP	**80.70%**

TABLE 6.2: Supervised Learning Results On Amazon dataset

In addition, as stated in section 6.2.3, introducing a verified purchase feature, which indicates whether the reviewer purchased the product, significantly improved the model's accuracy. This finding is consistent with previous research conducted by Mukherjee et al.[8], Birim et al.[12], and Barbado et al.[4] This highlights the importance of considering reviewer-centric features, such as verified purchase, in accurately classifying fake reviews.

Chapter 7

Discussion and Conclusions

This chapter will delve deeper into the results presented in the evaluation section of the project and reflect on the findings as a whole. The evaluation section provided us with quantitative and qualitative data that allowed us to assess the effectiveness and success of the project. We will analyze and discuss the data to draw clear conclusions on both the quantitative and qualitative aspects of the project. Furthermore, this chapter will examine how the results align with the initial objectives and hypotheses of the project. Finally, this chapter will present a conclusive summary of the project, highlighting its contributions, limitations, and future implications.

7.1 Project Review

The project implementation was divided into three phases, each with its objectives and results. In the first phase, natural language processing was performed on a real world Yelp dataset and trained supervised learning models to classify a review as real or fake. In the second phase, the attempt to label an unlabelled dataset using a clustering technique called topic modelling was unsuccessful. This led to phase three, where the results of topic modelling was incorporated into a supervised model for creating a more accurate machine learning model which learns from the topics in a corpus

In Phase 1, Naive Bayes, AdaBoost, and SVM algorithms were used to perform supervised learning using a Yelp dataset containing labeled real and fake reviews. The primary feature utilized in the analysis was the content of the reviews, and the results of the experiments showed that Naive Bayes achieved an accuracy of 67% and an F1 score of 78%, while the SVM algorithm achieved an accuracy of 62.70% and an F1 score of 78.7%. AdaBoost achieved an accuracy of 67.8% and an F1 score of 80.3%.

It was also observed that fake reviews had a much higher probability of having prepositions, while real reviews had a much higher probability of using conjunctions. In addition, real reviews had a significantly higher word count than fake reviews, consistent with previous research on the topic. **These observations suggest that fake reviews may use simpler language structures to avoid detection, while real reviews may use more complex language structures because they are written by genuine customers who want to provide detailed and informative feedback**.

In Phase 2, an attempt to cluster fake reviews using topic modelling was made. However, It became apparent that this assumption was flawed, as topic modelling aims to identify topics by grouping words that are more likely to occur together and does not inherently distinguish between real and fake reviews. Therefore, the initial approach was unsuccessful.

However, an attempt to answer the research question was still made, the question being - **Can topic modelling be used to detect fake reviews?**. To do this, phase 3 was introduced. In Phase 3, the density of a topic within a review was used as a feature. The results showed that while the accuracies of the Naive Bayes (NB), Random Forest (RF), and AdaBoost models were not affected, the accuracy of the Support Vector Machine (SVM) model increased. These observations were found to hold true for both the Amazon and Yelp datasets.

This suggests that incorporating topic modelling alongside models that are more sensitive to feature scaling, such as SVM, may lead to improved accuracy in detecting fake reviews. Specifically, using topic density as a feature may allow for more nuanced analysis of review content beyond standard text-based features, leading to greater accuracy in detecting fake reviews.

7.2 Conclusion

Based on the results of the study, the main conclusions were that **using only review centric features, such as the text content of the reviews, could detect fake reviews with moderate accuracy. However, the addition of reviewer-based features such as the number of reviews written by the reviewer, whether the purchase was verified, and the rating of the reviewer could significantly increase the accuracy of the models.** Random forest was found to be the best model for detecting fake reviews. Additionally, the study found that adding topic modelling as a feature is useful, but only when using a model more sensitive to feature scaling, such as SVM.

These findings suggest that it is **crucial for e-commerce companies to introduce reviewer-based features such as helpful review, likes, dislikes,verified reviever, verified purchaser etc. to their users**, as it can help maintain the integrity of their platforms by identifying and removing fake reviews. Furthermore, releasing this data to the public can enable researchers to conduct more studies using real-life data, leading to better and more accurate models in the future. Overall, the findings of this project provide valuable insights into the complex issue of fake reviews and offer practical recommendations for e-commerce companies to combat this problem, the study demonstrates the effectiveness of machine learning techniques in identifying fake reviews and highlights the importance of incorporating reviewer-based features in future studies.

7.3 Future Work

The current study has made significant contributions in identifying the most effective review centric features for training machine learning models to detect fake reviews. The study incorporated various review centric features such as review length, part of speech analysis, topic density, sentiment score, and review text. However, the results showed that combining all these features only produced a moderate level of accuracy. It was also observed that incorporating reviewer centric features, such as the number of reviews written by the reviewer, verified purchase, and reviewer rating, could significantly increase the accuracy of the model. Based on these findings, the most appropriate future research would be to experiment with different combinations of both review centric and reviewer centric features to identify the best combination for detecting fake reviews using machine learning. This could involve exploring different machine learning models or even deep learning techniques to improve the accuracy of the detection process. Additionally, extending the study to cover other review platforms and languages could further improve the accuracy and generalizability of the findings.

Bibliography

[1] T. Chen, P. Samaranayake, X. Cen, M. Qi, and Y.-C. Lan, "The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study," *Frontiers in Psychology*, vol. 13, 2022.

[2] M. Luca, "Reviews, reputation, and revenue: The case of yelp. com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, no. 12-016, 2016.

[3] N. Jindal and B. Liu, "Review spam detection. acm," *City*, 2007.

[4] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[5] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 219–230.

[6] E. D. Wahyuni and A. Djunaidy, "Fake review detection from a product review using modified method of iterative computation framework," in *MATEC web of conferences*, vol. 58. EDP Sciences, 2016, p. 03003.

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

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

[9] D. H. Fusilier, M. Montes-y Gómez, P. Rosso, and R. G. Cabrera, "Detecting positive and negative deceptive opinions using pu-learning," *Information processing & management*, vol. 51, no. 4, pp. 433–443, 2015.

[10] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[11] 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)*. IEEE, 2016, pp. 317–320.

[12] Ş. Ö. Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, S. Kumar, and Y. Kazancoglu, "Detecting fake reviews through topic modelling," *Journal of Business Research*, vol. 149, pp. 884–900, 2022.

[13] [Online]. Available: https://www.fakespot.com/

[14] [Online]. Available: https://pasabi.com/

[15] Z. Xu, B. Chen, S. Zhou, W. Chang, X. Ji, C. Wei, and W. Hou, "A text-driven aircraft fault diagnosis model based on a word2vec and priori-knowledge convolutional neural network," *Aerospace*, vol. 8, no. 4, p. 112, 2021.

[16] [Online]. Available: https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

[17] K. M. Leung, "Naive bayesian classifier," *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, vol. 2007, pp. 123–156, 2007.

[18] V. K. S. Reddy, "Stock market prediction using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 10, pp. 1033–1035, 2018.

[19] E. Fitzpatrick, J. Bachenko, and T. Fornaciari, "Automatic detection of verbal deception," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 3, pp. 1–119, 2015.

[20] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[21] S. Zhao, Z. Xu, L. Liu, M. Guo, and J. Yun, "Towards accurate deceptive opinions detection based on word order-preserving cnn," *Mathematical Problems in Engineering*, vol. 2018, 2018.

[22] S. Banerjee, A. Y. Chua, and J.-J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proceedings of the 9th international conference on Ubiquitous Information Management and Communication*, 2015, pp. 1–7.

[23] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 456–481, 2016.

[24] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering: An Open Access Journal*, vol. 2, no. 1, pp. 602–609, 2014.

[25] S. T. Dumais *et al.*, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.

[26] T. Hofmann, "Probabilistic latent semantic analysis," *arXiv preprint arXiv:1301.6705*, 2013.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[28] V. Tomaselli, G. G. Cantone, and V. Mazzeo, "The polarising effect of review bomb," *arXiv preprint arXiv:2104.01140*, 2021.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[30] M. Glinz, "On non-functional requirements," in *15th IEEE international requirements engineering conference (RE 2007)*. IEEE, 2007, pp. 21–26.

[31] F.-L. Li, J. Horkoff, J. Mylopoulos, R. S. Guizzardi, G. Guizzardi, A. Borgida, and L. Liu, "Non-functional requirements as qualities, with a spice of ontology," in *2014 IEEE 22nd International Requirements Engineering Conference (RE)*. IEEE, 2014, pp. 293–302.

[32] J. Horkoff, "Non-functional requirements for machine learning: Challenges and new directions," in *2019 IEEE 27th international requirements engineering conference (RE)*. IEEE, 2019, pp. 386–391.

[33] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[34] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[35] X. Xie, J. W. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, "Testing and validating machine learning classifiers by metamorphic testing," *Journal of Systems and Software*, vol. 84, no. 4, pp. 544–558, 2011.

[36] Y. Sun, S. Todorovic, and J. Li, "Reducing the overfitting of adaboost by controlling its data distribution skewness," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 20, no. 07, pp. 1093–1116, 2006.

[37] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction.* IEEE, 2013, pp. 245–251.

[38] S. Koltcov, O. Koltsova, and S. Nikolenko, "Latent dirichlet allocation: stability and applications to studies of user-generated content," in *Proceedings of the 2014 ACM conference on Web science*, 2014, pp. 161–165.

Appendix A

Code Snippets

Put appendix material in this section e.g. code snippets

USE THE APPENDICES

Appendix B

Wireframe Models