Krishna Patel

Comparative Analysis of Machine Learning Models for Detecting Fake Reviews on Amazon

Metropolia University of Applied Sciences

Master of Engineering

Information Technology

Master’s Thesis

1 January 2025

PREFACE

As an android application developer working on this topic was quite challenging but more rewarding. During coursework of Artificial Intelligence with Python I found interest in AI and decided to go deep diving in the AI era.

In the course of this study, I examined the use of machine learning models for detecting fake reviews on Amazon. Through the comparison of different machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forest and Gradient Boosting, I hope to find out which of the models will prove to be the most efficient when it comes to discriminating between genuine and fake reviews.

With this research, I aim to make a contribution to this field and apply the machine learning techniques to find the deceptive patterns and increase the review authenticity. Overall, I want to help make online reviews trustworthy and give businesses a set of tools to fight the fraud with the reviews.

I am very grateful to my thesis instructure Toni Spännäri, who provided guidance, support, and thoughtful feedback throughout this work and encouraged me to complete within a certain deadline. Many thanks to Ville Jääskeläinen for helping me finalize the thesis topic.

Finally, my heartiest thanks from the bottom of my heart to my family members, especially my daughter(Niyati Patel), my husband, parents and in-laws for their support, patience and motivation throughout this study. Their understanding during the late nights and busy weekends helped me finish this thesis.

Espoo, 25 May 2025
Krishna Patel

Abstract

Author:	Krishna Patel
Title:	Comparative Analysis of Machine Learning Models for Detecting Fake Reviews on Amazon
Number of Pages:	56 pages + 9 appendices
Date:	1 January 2025
Degree:	Master of Engineering
Degree Programme:	Information Technology
Professional Major:	Networking and Services
Supervisor:	Toni Spännäri, Senior Lecturer

This research is concerned with the efficiency of machine learning models when it comes to detecting fake reviews on Amazon. Since or because of the rapid growth of e-commerce, online reviews have become important in determinations of consumer decisions. But the growing trend of fake reviews erodes the trust of the consumers and alters the behavior of the market. There are various evaluations of machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting among others to establish the most effective and reliable model for detection of fake reviews.

This study uses a publicly available dataset of Amazon product reviews identified as either genuine or fake, which contains text data as well as metadata on reviewers. Before training the models, methods of data preprocessing are applied, including the text cleaning, tokenization, and feature extraction. Performance evaluation is done based on the metrics of accuracy, precision, recall, and F1 score. Results reveal that the ensemble methods such as Random Forest, and Gradient Boosting classifiers perform better than other models in terms of recall as well as overall classification performance. The study identifies the issues with processing imbalanced datasets and points to its importance to pay attention to model transparency and interpretability. Lastly, the research offers recommendations to e-commerce platforms in order to increase the review credibility and safeguard consumer trust.

Keywords: ML, Amazon, review detection, LR, SVM, RF, GB, E-commerce, text categorization, data preprocessing, accuracy, recall, precision, F1 score, model assessment, dataset, genuineness of reviews, trust of the consumers.

Contents

List of Abbreviations

AUC - Area Under the Curve

EDA - Exploratory Data Analysis

F1 Score - F-Measure Score

GB - Gradient Boosting

KNN - K-Nearest Neighbors

LR - Logistic Regression

ML - Machine Learning

NLP - Natural Language Processing

PII - Personally Identifiable Information

RF - Random Forest

RFE - Recursive Feature Elimination

ROC - Receiver Operating Characteristic

SMOTE - Synthetic Minority Over-sampling Technique

SVM - Support Vector Machines

TF-IDF - Term Frequency-Inverse Document Frequency

1 Introduction

1.1 Background of Online Reviews

Today online reviews have become an integral part of the e-commerce ecosystem that essentially define consumer purchasing behaviour. The popularity of e-commerce platforms like Amazon has shifted the customers towards digital reviews, to read and observe the quality of a product, to check the different alternatives around it, and ultimately deciding wisely. These reviews become digital word of mouth which gives an insight from other buyers and builds trust that’s otherwise in particular transactions. As per several studies, there's nothing that quite affects buying until and unless a review is published, either a positive or negative one influences the buying choices considerably.

Whilst the open style of review systems is also a source of the spread of fake or manipulated reviews (Pfänder and Altay, 2025). While bots, incentivized users and even competitors may produce these reviews with the aim of deceiving potential buyers, these reviews are not verified. Fakery, whether in the form of absurdly high ratings for a product or the defaming of a seller’s reputation at the cost of an artificially bad rating, is capable of distorting the play of the market. Not only does it mislead consumers, but also e-commerce platforms lose credibility, when the false information on e-commerce platforms is so easy to copy and steal. Consequently, it has become a crucial challenge for businesses and researchers to detect and filter fake reviews. Machine learning has already made numerous progress in identifying deceptive patterns, authenticating user generated contents, which will ultimately ensure that the consumer trust in online marketplaces is maintained.

1.2 Research Problem

Fake reviews, on the other hand, refers to deceptive or fraudulent evaluations that are posted online with the intention to trick customers and sway opinion about the product or the firm. Usually these reviews are written by people or even automated bots, and either to promote a product unfairly (positive fake reviews), or damage a competitor unfairly (negative fake reviews) (Christiaens, 2025). A genuine review is based on what a user has experienced in real life, but fake reviews are nothing but misleading and were used for a particular agenda, most of which are economic or competitive.

In e-commerce platforms, customer reviews have become increasingly important and both have become opportunities and vulnerabilities. Reviews are a form of social proof that influence a business's product visibility and sales ranking. Reviews are therefore critical for consumers to decide product quality and what to purchase. But fake reviews undermine this trust and have become the basis for distorted market behaviour, misallocated consumer spending and harm to honest sellers reputation. From an economic point of view, fake reviews can help reduce consumer trust, increase return rates and generally decrease the integrity of digital marketplaces such as Amazon.

While there has been increasing awareness of the issue, no simple and straightforward solution [to fake reviews] exists. Due to the ability of fake reviews to duplicate genuine patterns in language and behaviour, traditional rule based filtering techniques tend to fail (Jaoua et al. 2025). In addition, the volume of reviews is high and the sophistication of these tactics precludes manual moderation. Thus this study can argue that machine learning provides a promising way to conceive of scalable and adaptive approaches to analyze the linguistic cues, reviewer's behavior and metadata to identify suspicious content. Nevertheless, to select this most effective model, this study has to evaluate all algorithms and their skill in generalisation over the different datasets of review.
The overall goal trying to achieve in this study is doing a comparative analysis of machine learning models to determine the most accurate and reliable

machine learning models to find fake reviews on Amazon. The work will extend the line of work on enhancing review authenticity and protecting trusts of customers in e-commerce.

1.3 Research Objectives

The aim here is to evaluate and compare the performance of different ML models in accurate detection of fake reviews on Amazon platform, thereby improving review authenticity and trust of e-commerce platforms.

Research Objectives

- To investigate the characteristics and patterns typical to fake reviews.
- To preprocess and analyse Amazon review datasets in order to train machine learning models.
- To implement multiple algorithms of machine learning for fake review detection.
- To Utilise metrics such as accuracy, precision, recall and F1 score will be used to evaluate model performance.
- To deploy in reality in e-commerce platforms, this thesis recommends the most appropriate model.

1.4 Scope and Limitations

Specifically, this study centres its analysis on the task of detecting fake reviews within the extensive and influential context of Amazon, the largest and practically most celebrated e-commerce platform in the world. This scope involves use and comparison of different ML models namely, Logistic Regression, Support Vector Machines (SVM), Random forest & Gradient Boosting to predict deceptive patterns (Farsi and Chowdhury, 2025). The work consists of data preprocessing, features extraction, model and metrics like accuracy, precision, recall, and F1-score evaluation. This study seeks to find which is the best model to distinguish between the fake reviews that preserve the generalizability and efficiency.

The research will use at least one of the publicly available datasets such as the ones from Kaggle and others academic repositories including labelled Amazon product reviews. For instance, these datasets usually have info such as review text, ratings, reviewer profiles and timestamps that could be utilised to construct reliable classification models. The focus is primarily on technical review fraud, with data based and experimental work(domain only) in the study and very minimal psychological or behavioural motivations for what drives review fraud. However, many limitations must be noted. For example, the availability and the quality of labelled data may be a hurdle for the data, as most of the real world diversity of review is not pre identified as fake or genuine. That may also impact the models' generalizability to newer or more nuanced forms of fake reviews (Salminen et al. 2025), i.e., the size and diversity of the dataset may be limited by this. Further, the study does not include deep learning approaches, including recurrent neural networks and transformer-based models, as they would take a lot of time and computing resources. More specifically, the limit of the research is focusing on the textual and behavioural analysis of the reviews and not given the attention of the other contextual factors such as product categories, market trend, and the reviewer sentiment over time.

This limitation aside, the study sheds light into the level of effectiveness of the machine learning techniques in improving e-commerce user generated content integrity.

2 Literature Review

2.1 Evolution of Review Analysis

A huge progress toward the change of consumer behavior and decision while shopping for goods on the internet was marked by the rise of online reviews. It evolved as user generated comments on product pages or forums in the early 2000s, and it then gained strength as a form of online word of mouth (Singh et al. 2024). As central features to expand the opinion and experience sharing by consumers, e-commerce giants such as Amazon, Yelp and TripAdvisor started integrating review systems through which they could express their opinion and share their experiences. Since reviewers' influence on customer trust and product sales increased, the legitimacy of those reviews became an issue.

The sentiment classification was mainly focused in earlier phases of review analysis. Also, researchers used such basic Natural Language Processing (NLP) techniques as lexicon-based approaches and rule based classifiers to score positive or negative opinions (Gunasekaran, 2023). While these methods were sophisticated enough to distinguish between deceptive and legitimate content, they were, however, quite simple. First, there were some early studies that started noticing the review spamming and proposed supervised models that would detect anomalies by language patterns and metadata.

When the research community became aware of the fake reviews problem, they started to investigate review authenticity. That shift came about due to the fact that it was realized not all sentimentally positive or negative reviews are true. Such proliferation of fake reviews (incentives or by bots) was a significant threat to consumer trust and to integrity of review systems.

Successful approaches that followed to detect review authenticity entailed machine learning models capable of learning from labeled data. The initial widely used algorithms include Logistic Regression, Naïve Bayes, and Support Vector Machines (SVMs). These models made use of such features as review

length, frequency of review posting, and reviewer behavior, as well as linguistic cues. While they had some efficacy to some degree, they struggled hugely to be flexible enough to different datasets and different product categories.

During the past decades, a number of ensemble models such as Random Forests and Gradient Boosting have been shown to enhance the prediction accuracy of many learners (Rane et al. 2024). The effects of interactions between variables can be taken into account by these methods, and such complexity is less likely to result in overfitting. In the meantime, it examined reviewer credibility, history and consistency to supplement authenticity detection (Abdulqader et al. 2022).

Review analysis is one of a series of evolution of a method from its development to the final incarnation. The complexity and creativity of deceptive reviews have already made great strides in traditional models, yet they are increasingly sophisticated methods required. This consists of deep learning as well as many more future exploration for contextual modeling. But the literature concerning this is nonetheless strong and makes for a fruitful comparative study, for example to evaluate how different machine learning models perform to classify real and fake reviews on Amazon and other such platforms.

Due to the increasing fake reviews manipulation on e-commerce platforms, detecting fake reviews has turned out to be a critical research area. And through time, various techniques were developed such as manual detection, automated machine learning and advanced deep learning models. The way of modeling discussed above has its own strengths and weaknesses based on the complexity, scalability, and precision in real world applications.

2.2 Existing Fake Review Detection Techniques

The earliest type of fighting with deceptive content was manual selection of a fake review. It involves having human moderators or analysts reading and evaluating reviews manually and based on criteria and features such as tone,

specificity, and review history. The latter typically are used in moderation systems where a suspicious review is flagged by users or algorithms and reviewed by a human team.

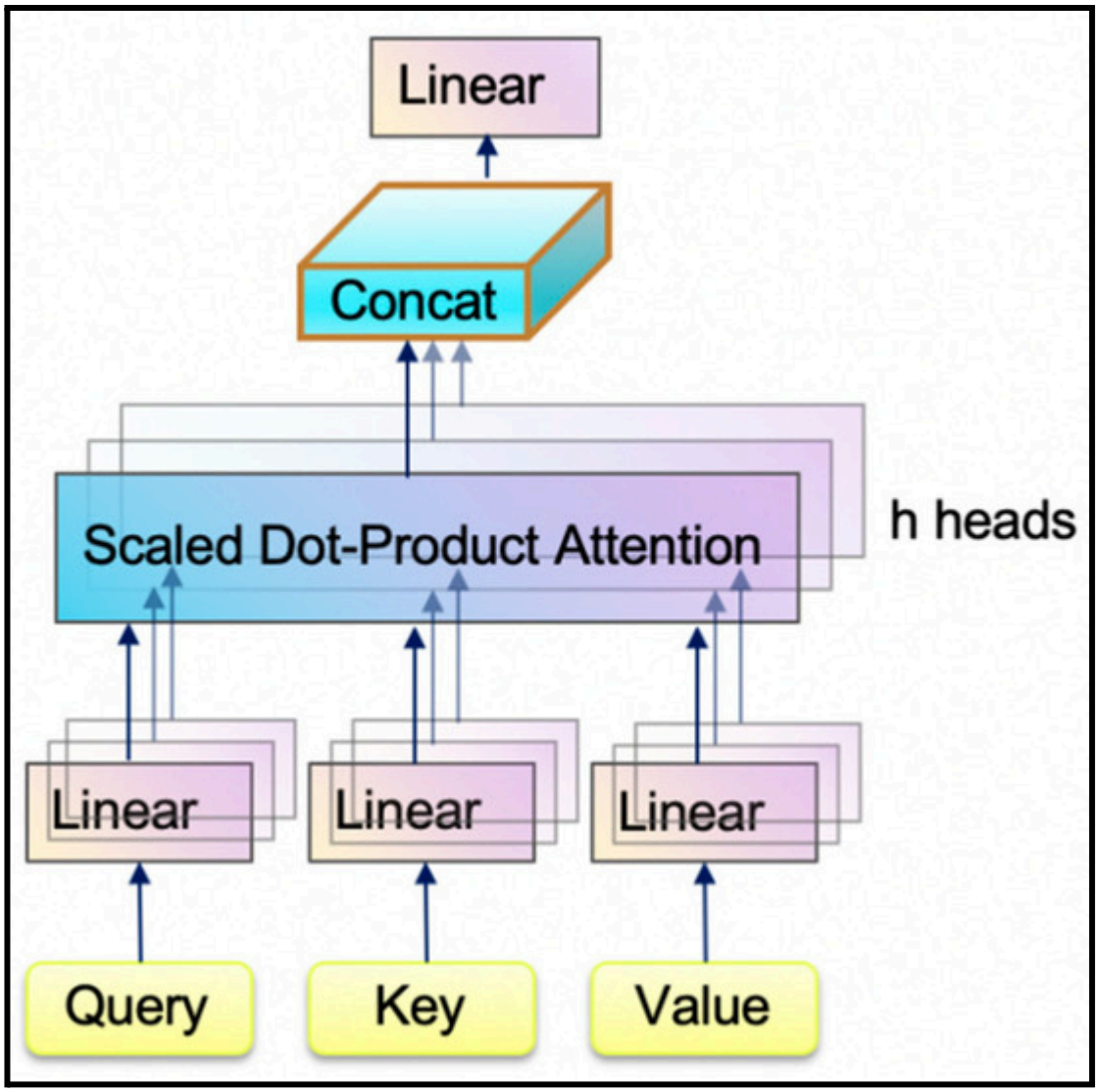


Figure 1: Fake Review Detection

(Source : mdpi.com, 2024)

Manual Detection Methods

The first attempt to deal with deceptive content was manual detection of fake reviews. It consists of human moderators or analysts manually reading through reviews and assessing the reviews in regards to their tone, specificity or any background information. Generally, these are employed in a moderation

system, where the suspicious review is tagged by users or algorithms and then a human team evaluates it.

The objective of manual detection is inherently limited in both the scalability and objectivity. Human reviewers may be able to pick up on a few subtle nuances in language, such as sarcasm or overly generic praise, but they may differ and be too slow, especially in the amount of daily reviews that popular online platforms such as Amazon are generating. Furthermore, manual systems are faulty and prone to bias, poor and unsuitable for real time or large scale review analysis. It has been largely supplanted or substituted by automated techniques, so manual detection is still used in applications on the small scale or as a verification step (Farsi and Chowdhury, 2025).

Traditional Machine Learning Approaches

A high amount of automation in fake review detection is due to the emergence of traditional machine learning models. They construct these models, which learn patterns from pre-labelled datasets for which reviews are marked as genuine or fake. Applications of this type of algorithm include Logistic Regression, Support Vector Machine (SVM), Decision Tree, Naïve Bayes, Random Forest and Gradient Boosting Machine, which are the widely used algorithms in this category.

They are based on feature engineering, that is, on manual selection and transformation of attributes that describe a review in input variables. Commonly used features include:

Text-based features: word frequency, sentiment polarity, part-of-speech tags.
Reviewer behavior: how recently they reviewed, what was the average rating given by them, how many reviews have they written within the scope of our criteria.
Metadata: review length, presence of verified purchase badge, timestamp patterns.

To name a few, Logistic Regression and Naïve Bayes have proved to be fruitful in the simpler and most basic binary classification, based on textual features. Perhaps, SVM performs well in high dimensional space with great performance in case of kernel trick to detect the separation of non-linear spaces (Khan et al. 2023). Unlike Decision Tree, Random Forests and Gradient Boosting make use of multiple decision trees to achieve better classification performance as well as to mitigate overfitting.

Traditional ML models work decently but are hard to interpret, to some extent, and to deploy. Moreover, these models usually heavily rely on extracting good quality features as well as having labelled training data, which may or may not be present, and may not be consistent across platforms.

Deep Learning Techniques

During the era of large scale data and developed computational power, deep learning methods have become attractive reagents to the conventional ML for fake reviews detection. Raw data can be learnt with deep learning models with manual pre-processing, where the latter is minimal (Gunasekaran, 2023).

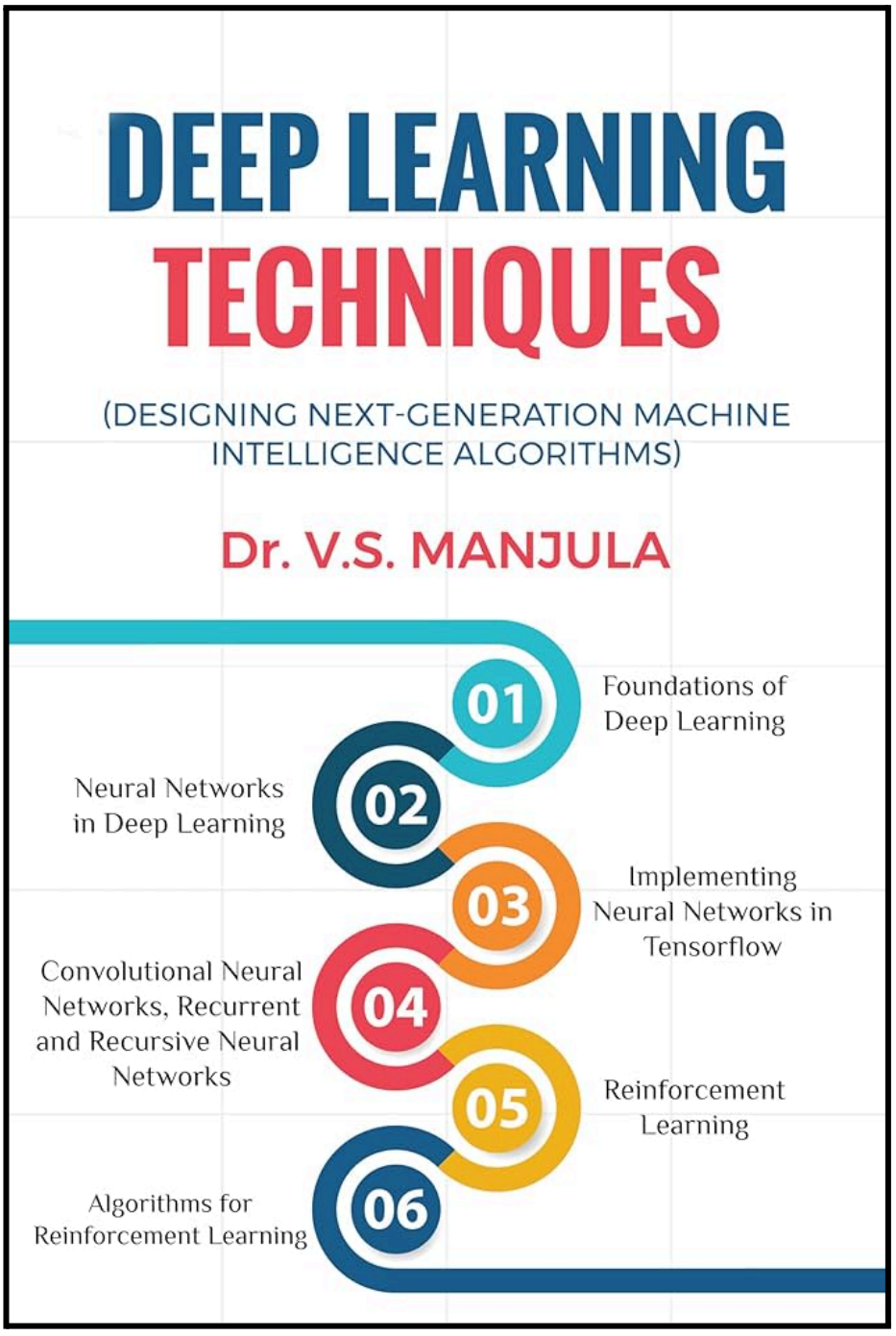


Figure 2: Deep Learning Techniques

(Source: www.amazon.in, 2024)

The widely used deep learning approaches include some examples such as:

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): These are models that can process the sequential data and are quite

extensively used to capture sequential linguistic patterns in review texts. Text can be forgotten depending on its position in the sentence, making a bag of words models not the best way to understand the context and flow of text.

Convolutional Neural Networks (CNNs): Originally, CNNs had been designed for image processing by treating text as a matrix of word embeddings for NLP tasks. When searching local patterns such as particular phrases or word combinations that frequently exhibit in fake reviews, they are effective.

Models Based on Transformers: Like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), the latter ones have changed the course of NLP by allowing deep contextual understanding of any text (Yenduri et al. 2023). For example, BERT understands how words relate to each other in forward and backwards directions well, which makes it possible to observe the small signs of deception and exaggeration in reviews.

In many cases, these deep learning models outperform the traditional techniques and in particular in handling large, complex and noisy datasets. Additionally, they can also adapt better to many different domains and languages. Nevertheless, they are very computer intensive and need large amounts of labelled training data. This can also be a concern if one is using “herding” algorithms to optimize e-commerce platforms since their “black box” nature sometimes can make them difficult to interpret and validate.

Finally, this summarises that fake review detection has moved from a simple manual method to machine to deep learning models. Manual approaches are useful in small scale and for high risk content but it is unimaginable at scale. By incorporating scalability to detection, traditional machine learning models can be used, but features need to be engineered first and more data labelled. However, state of the art performance is achieved by deep learning techniques, especially transformers but with computational and interpretability challenges (Khan et al. 2023). This evolution shows the applicability of evolving online

deception to complexity and scale increase and the desire for an adaptive, accurate and efficient system of detection. This paper compares some of these traditional machine learning methods and which one provides a more reliable solution from the point of view of detecting the fake reviews on Amazon.

2.3 Machine Learning Models in Review Analysis

In fact, machine learning is a very useful tool for review analysis, especially the fake review detection, sentiment classification and opinion mining tasks. Two typical categories of approaches are given to be based on supervised learning and unsupervised learning, which both have their strengths in dealing with different facets of review data.

Supervised Learning Methods

Supervised learning refers to training a model on a dataset which has been labelled and to know its result (for example, that the picture is fake or genuine, positive or negative). It is one of the widely used approaches for fake review detection and sentiment analysis. My examples encompass Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, Neural Networks.

The features of the model are derived from review analysis that includes the word frequency, n-grams, sentiment score as well as the reviewer metadata (e.g., review length, time of posting and verified purchase). One example is that a model could be trained to separate from fake reviews using previously labeled examples by, for instance, learning with linguistic patterns or suspicious reviewer behaviour (Abdulqader and Alsaawy, 2022).

Supervised learning is strong at prediction, as it can be quite accurate, provided that it has plenty of good high data quality labelled data. But it permits to fine tune models capable of learning intricate decision boundaries. Nevertheless, dependency on availability of annotated datasets, especially datasets in niche

or emerging domains, is very high and production of such datasets are expensive and time consuming.

Unsupervised Learning Approaches

While unsupervised learning does not require any labelled data, supervised learning does. It does not suggest patterns, structures or groupings instead, but rather finds hidden patterns within the review data. K means clustering, DBSCAN, Latent Dirichlet Allocation (LDA) and the like.

With regard to review analysis, unsupervised models allow us to group similar reviews, find out emerging topics or sentiment, and detect outlier behavior which hints for spam or fraudulent behavior. For example, clustering can be used to identify, for example, groups of reviewers who consistently give too positive feedback for products that are rated less highly, suggesting that it could be manipulated (Deshai and Rao, 2022).

Unsupervised methods are best suited when the data are unlabelled or if one is simply performing an exploratory analysis or initial screening. But they're often more hand interpreted and their results are less precise or may not constitute a direct classification output.

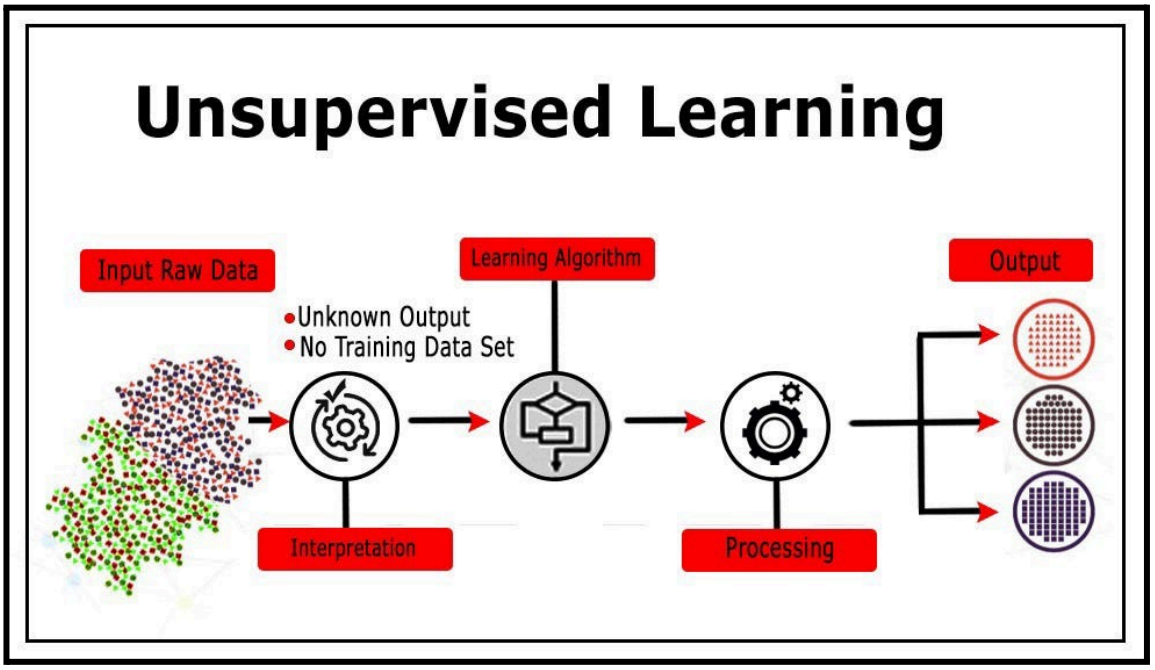


Figure 3: Unsupervised Learning

(Source: techntales.medium.com, 2024)

Hence, supervising and unsupervised learning are significant in reviewing. Supervised methods are good used for tasks needing high accuracy, but unsupervised ones provide practical wisdom especially in the early stage or data sparse environments.

Hybrid detection techniques

Hybrid detection techniques use a set of machine learning models or methodologies for better accuracy and efficiency in fake review detection (Deshai and Rao, 2022). The objective is to use the lot to mitigate the weakness of the one, while taking advantage of the strength of the other. When applied to fake review detection, hybrid models usually adopt supervised and unsupervised learning or a fusion of the traditional machine learning models with the deep learning ones.

The common hybrid approach is feature engineering with the advanced machine learning algorithm. As an example, Random Forest, Support Vector Machines (SVM) or Gradient Boosting can be used for classification with text based features, for example word frequency, sentiment analysis and linguistics patterns extracted. This combination makes the model good at handling both structured (e.g., ratings, history of the reviewer) along with unstructured data (e.g., text review content) that helps detect sophisticated fake reviews.

Hybrid techniques, other than those involving the use of meta models, make use of ensemble techniques like the use of multiple models that have been trained in isolation and then the output of the various models are combined in the final prediction. It helps make the robustness and reliability of the detection system better by means of example, combining the predictions of SVM and Random Forest or using a voting mechanism (Hossain and Islam, 2023).

In addition to that, it can be with a combination of both supervised and unsupervised learning methods. Firstly, cluster models can be unsupervised, with the potential suspicious reviews being identified, and then supervised classifiers used to confirm them as such. The result of this two step process is that it will increase the precision of detecting fake reviews while reducing false positives.

Overall, a hybrid detection technique, as presented above, provides a promising path to improve scalability, accuracy and adaptability of fake review detection systems.

2.4 Theoretical Framework

2.4.1 Key theories underlying fake review detection

The field of fake review detection is interdisciplinary and it uses multiple theoretical frameworks to understand and tackle the problems of identifying fake content in Online Reviews. The key theories on which fake reviews are detected are language patterns, behavior analysis and the larger context of trust in online systems.

Linguistic Theory

According to theories of language use and deception, linguistic analysis has a central role to play in the fake review detection task (Abdulqader et al., 2022). According to the Deceptive Speech Act Theory, deceptive communication, such as fake reviews, generally corresponds to the less frequent patterns of truthful communication. That includes steeping in overly positive or negative language, picking out weird term uses, or tone… it doesn't match. Specific linguistic cues such as excessive use of positive adjectives, superlatives and generic phrases are exhibited by the reviewers of the fake content. These linguistic features are used by machine learning models as a base for authentic fake review classification.

Behavioral Trust Theory

Behavioral Trust Theory focuses on trust being developed and reinforced since trusted actions or interaction are repeated in a reliable manner (Xie et al., 2024). Users tend to trust reviews posted by reviewers who have been reliable and verified in the case of online reviews. But fake reviews frequently erode this trust because they come from dubious and fake sources, for example, forums or bots or competitors. This theory informs us why there is a need for systems that detect deviating behaviors from expected performance, which is a leading sign of fraudulent activity.

Signal Detection Theory

Another theoretical approach for fake review detection is Signal Detection Theory (Batailler et al., 2022). According to this theory, the model's decision-making ability to distinguish between truthful and deceptive reviews can be identified as a separating between “signal” (realized reviews) and “noise” (fake reviews). This is a probabilistic reasoning based theory for rating a review as fake based on the review content, meta data, user trace pattern.
Social Proof Theory

It should be noted that Social Proof Theory asserts that people make the majority of their decisions based on the actions or opinions of other people, especially when they’re unsure. And that theory goes that consumers have a lot of trust in reviews in their case of online reviews. However, as fake reviews will always contradict truth, it damages this social proof mechanism. It is critical to be able to detect fake reviews — for consumers, this means that they are allowed to rely upon authentic feedback to make a purchasing decision, to maintain an effective social proof in the digital marketplaces.

These theories are combined to provide a general framework of complexity of the fake review detection problems and choice of suitable machine learning models to solve this problem.

2.4.2 Conceptual model of review authenticity

A review authenticity conceptual model is a structured method of understanding various relating factors and processes of authenticity determination of an online review. Additionally, this integrates various scope on the basis of the user behavior and content features, as well as metadata to assess the reliability of reviews from digital marketplaces such as Amazon. The review content feature space that is the center of the model consists of words, sentiment, grammar, exaggerated language, within a phrase, etc. In general, any review which is authentic usually contains diverse and nuanced language as compared to a fake review which has a trend of overly simplistic, generic, overly positive, or overwhelmingly negative tone. In this dimension, it is an important tool for sentiment analysis, because real reviews, or at least the best ones, can have something positive and negative balanced, while fake reviews have one aspect. The other crucial piece is the behaviour of the user: how often do they review submissions, what was the history of the reviewer, and what happens when interacting with the platform (Deshai and Rao, 2022). Usually fake reviewers will have a history of imbalance and subjective feedback while their trusted reviewers will have a more consistent history of balanced and thoughtful feedback. Reviewing the profiles of the reviewers, for example, verified purchase status, social proof indicators, etc., can also help to assess authenticity.

The conceptual model also includes metadata such as the date of review, product categories and the amount of reviews. Also, abnormal consumer behavior such as reviews appearing out of nowhere or abnormally high are suspicious reviews. Considering all these factors, the model provides a general examination of the content authenticity, and builds a foundation to generate the ML algorithms to detect such fraudulent material.

3 Research Methodology

3.1 Research Design

In this context, the research took a comparative study approach to assess and to identify the ideal machine learning model to use in detecting fake reviews on Amazon (Tabany and Gueffal, 2024). Systematic evaluation of training and testing multiple models under all similar conditions with a comparative analysis between them will give the idea of the difference in performance, strengths, and weaknesses of that model. By using this approach one can be assured objective with which algorithm generalizes better across unseen review data, that is critical for real world deployment in changing e-commerce environments. Based on the background in text classification tasks and use of fake review detection studies, Logistic Regression, Random Forest, Support Vector Machines (SVM) were selected as machine learning models to be used (Elmogy et al., 2021). i.e logistic regression is simple and easy. Through the ensemble learning, Random Forest is provided with robustness against overfitting. Since text based features are high dimensional, SVM is good at high dimensional feature spaces. They are able to capture non-linear relationships and complex patterns which in turn gives more prediction performance.

A different family of learning techniques (linear, tree based, kernel based, deep learning) is represented in each model, hence there is a holistic comparison across different complexities and learning paradigms. This comparative structure is crucial for recommending a sensible and honest model that could be used to observe deliberately creating reviews on Amazon’s gigantic virtual showcase.

3.2 Data Collection

This study uses fake reviews dataset pulled from publicly accessible repositories such as Kaggle and academic datasets to create the dataset used. The exact details of these datasets are labeled instances of genuine and fake

reviews (including their review text, ratings scores, reviewer metadata (e.g. review times and review timestamps, verified purchase status), and reviewer profiles. Labels of datasets help the machine learning models to be trained with the clear definition of positive and deceptive examples in a supervised manner.
Thus, the data method involves datasets fetching, asserting the same labels to be in coherence, and maintaining a well balanced set of fake and real reviews to not bias the model. Before feeding your data into the machine learning algorithms, previous preprocessing techniques were necessary. The steps involved in this process were text cleaning (removing HTML tags, special characters, and stop words), tokenization, lemmatization, as well as encoding categorical metadata features.

The datasets were used that only make use of public datasets in order to not compromise user privacy while strictly following ethical considerations. The data was not collected nor processed for any personally identifiable information (PII). The responsible practices approach adopted by the study involves making it clear that model predictions are only used for academic research and not in an unauthorized or a commercial manner without consent. All the phases of a research take place within the principles of transparency, fairness and accountability.

3.3 Data Preprocessing

Before looking at building a machine learning model for fake review detection or sentiment analysis, data preprocessing is a vital step. Preprocessing the data handles the complexity and noisiness of the online review dataset so that the dataset is clean, structured for algorithmic interpretation.

1. Text Cleaning:
Review data is a raw medium, which is often noisy due to content such as HTML tags, emojis, special characters, URLs and stopwords (e.g., “and”, “the”, “is”). Regular expressions and natural language processing (NLP) tools remove these. In order to be consistent, also convert reviews to lowercase.

2. Tokenization & Lemmatization:

Individual words (tokens) are broken down into text. Lemmatization downsizes words to their root, that is destiny shape (e.g., “cheetahs” becomes “run”), decreasing the dimensionality but preserving meaning. This step is important because linguistic analysis is used to improve model performance.

3. Feature Extraction:

These texts are cleaned and tokenized in order to transform them into numerical features. Some of the common techniques are Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings (e.g. Word2Vec, GloVe). However, these features aid in recording the pattern, word significance and contextual meaning inside the reviews.

4. Metadata Handling:

It normalizes and encodes reviewer metadata e.g. the length of the review, the time of posting, whether it's a verified purchase, and how frequently a reviewer posts a review. They can further be divided into categorical and numerical features, the former being one hot encoded while the latter are standardized to common scale.

5. Handling Imbalanced Data:

Many of these fake review datasets are imbalanced — there tend to be fewer fake reviews than real ones. To account for this imbalance, account for it via techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or under sampling the majority class to prevent biased model performance.

6. Outlier Detection:

Anomaly detection is levied against anomalous review behaviors (e.g. an unusually high number of reviews in a short time period). If detected, data points like that can be removed, or otherwise flagged as suspicious for further supervised analysis.

Therefore, these preprocessing phases ensure that both the textual as well as behavioral aspects of reviews have been adequately represented, activating a strong fake review detection.

3.4 Machine Learning Models

For that, four machine learning models were run which include Logistic Regression, Random Forest, Support Vector machines (SVM), Neural Networks.

Binary classification tasks have a logistic regression as a baseline model. It predicts whether a review is fake or not by linear combination of a set of input features. It is extremely simple and it gives us some valuable interpretability and it is useful for quick baseline comparisons.

An ensemble technique, Random Forest is a method which creates multiple decision trees, and combines their generated decision to increase the classification accuracy and help prevent overfitting. It is also capable of handling feature interactions and non linear relationships which would be suitable for complex review data.

The goal of the Support Vector Machines (SVM) is to find the best hyperplane that differentiates between the fake and genuine reviews in a high dimensional space (Alsubari et al., 2022). The kernel trick solves the problem of handling non linear separations which is needed whenever those are not able to linearise the textual features.

Multilayer Perceptrons are a Neural Networks type which consists of layers of interlaced neurons capable of carrying intricate relations among features. On the other hand, neural networks tend to work well when confronted with a mixture of textual metadata features and behavioral data features.

Hyperparameters like C for SVM, number of trees for Random Forest, learning rate for Neural Networks, etc (Daviran et al., 2023). were tuned by grid search and cross validation to have maximum performance during training.

3.5 Feature Engineering

It was important to feature engineering that improved learning. As textual, metadata, and other review specific characteristics were considered three categories of features.

Textual features were extracted using some techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. It converted the raw review text into numerical vectors, keeping key terms and ignoring many common words. In addition, sentiment scores, bigram frequencies and syntactic markers, like parts of speech tags were included to capture fine linguistic patterns.

Feature metadata included things such as the length of the review, whether the reviewer was verified or not, when the review was performed, and the average rating behavior of the reviewer. The behavioral and temporal cues that characterize these reviewers are different between genuine and fake reviewers. It calculated review specific characteristics such as deviation of the individual review ratings from the product’s average rating. Suspicious activity can be detected in rating behavior anomalies.

In order to reduce dimensionality, remove redundant features and improve model generalization, it applied feature selection techniques such as Recursive Feature Elimination (RFE), chi-square tests, etc. This allowed overfitting to be avoided, the models to be interpretable, but not to lose critical information in the process, and only the most predictive attributes to be used in the training of machine learning models.

3.6 Model Evaluation Metrics

To rigorously evaluate the Machine Learning models, several evaluation metrics such as Accuracy, Precision, Recall, F1 Score and ROC AUC were used.

The accuracy is defined as the proportion of correctly classified reviews out of the total number of reviews including fake and real ones. Although applicable it may be misleading if the dataset has imbalance and another metric is needed.

For this, precision quantifies the ratio of true positives to all predicted positives. High precision in fake review detection guarantees that the majority of reviews flagged as fake actually are deceptive and hence leastens the ability to accuse legitimate reviewers for being fake.

Specifically, sensitivity (or recall) is the ratio of true positives to all actual positives. The model has a high recall, meaning that it reaches most of the fake reviews when aiming to clean the marketplace from deceitful content.
Harmonic means of precision and recall serves as a F1 Score which balances the trade off between any two metrics. In particular, this requires minimal cost for false positive and false negative when the cost of a mistake is equally important.

ROC-AUC (Receiver Operating Characteristic Area Under the Curve) is used to measure the model’s capacity to differentiate classes in the case of various threshold settings. The higher the AUC is, the better overall classiling ability our model will perform, and the better this metric is as a model performance comparison.

3.7 Ethical Considerations in Fake Review Detection

Fake reviews can be detected using machine learning techniques when dealing with user generated content but as it is ethical to consider it, it's crucial. The following gives the key ethical aspects about this research.

Data Privacy and User Consent:

In this study, publicly available Amazon reviews must be dealt with in a manner that complies with privacy laws such as GDPR. But, the data to be used for model training should not contain personally identifiable information (PII). All datasets should never be disclosed with respect to privacy of users, at the same time consent for using the data should always be clearly stated citing the purpose in academic research.

Transparency and Model Interpretability:

The lack of interpretability is considered one of the problems of machine learning models, especially of the deep learning models. To tackle this, Decision making should be transparent and explainable, providing clear reasoning behind why a review is classified as fake or real. Such enforcement helps to make the system accountable, as well as to promote trust in the system, which is quite critical in real world applications where business decisions lead to impact on businesses as well as on the customers.

Mitigating Bias:

The consequences of a bias to the machine learning model are that it can have unfair (not telling the whole truth) or inaccurate (not completely accurate) predictions such as labeling legitimate reviews as fake. To be fair, the research must make use of such methods to escape biases present at the data level. Using a balanced dataset and bias detecting algorithm can achieve this goal, which will make the model more equitable and reliable.

Finally, research must be guided by ethical standards where minimum data are privately, transparently, and fairly collected. This contributes to the integrity of the study and increases the trustworthiness of the machine learning models trained on detecting fake reviews in platforms such as Amazon.

Besides data privacy and user consent, transparency and model interpretability, and mitigate bias, there are a couple of additional ethical requirements for a fake review detection system to be responsibly developed and deployed.

False Positives & Reputational Harm:

Perhaps the most pressing of ethical issues is the possibility of false positives—wherein perfectly valid reviews were misidentified as being fake. The consequences of such errors can be significant: reputational damage to readers, suppression of genuine feedback, and damage to the credibility of genuine reviewers. To avoid the undue harm to users who are honest contributors to online platforms, high model precision is needed, and manual review mechanisms must be in place to flag handled content.

Accountability & Oversight:

With more complex fake review detection models, holding accountable gets even more important. In the last decade, developers, platform operators, and data scientists have yet to define clear roles and responsibilities for training, deploying, and monitoring models. Thus, it is necessary to develop ethical AI governance frameworks that prescribe how decision-making processes should be conducted, requirements for regular audits, and assessment of societal impact of the system (De Almeida et al. 2021). When there lacks accountability, the resulting outcomes could be harmful.

Continuous Monitoring & Adaptation:

Tactics of fake reviews change over time, and static models are likely to become quickly obsolete. The method is ethical towards the implementation which includes continuously monitoring model performance, retraining with new data, updating detection rules to keep being effective (Muskan and Wajid, 2024). However, by taking a proactive approach, the system is in a constant state of accuracy, with minimal unintended consequences, as it is reconfigured to reflect changes in online behavior patterns. In addition, user and human moderator feedback can help create a more ethical and responsive system.

These concerns are additional to the fundamental ethical principles of privacy, transparency and fairness and collectively contribute to developing trust in fake review detection technologies that align with the beliefs of the society and

preserve user rights.Ethical assessment must be maintained, particularly during deployment of AI in the context where people’s perception of an entity and consumer behavior could be impacted.

3.8 Chapter Summary

In the methodology chapter, I explain the systematic approach to find fake online reviews by several machine learning models. To start with, data is selected from e-commerce platforms specifically labeled datasets that bear both true and fake reviews. This aims to provide a scientific reproduction of a complex data preprocessing procedure, which involves text cleaning, tokenization, stop word removal, lemmatization and feature extraction (e.g., sentiment scores, linguistic patterns, reviewer metadata). Steps are taken to ensure high quality input into the model training.

This thesis uses both supervised and unsupervised machine learning methods, such as Logistic Regression, SVM, Random Forest, K-Means Clustering, and Latent Dirichlet Allocation (LDA). To improve accuracy, as well as to increase adaptivity, hybrid techniques combine these models. Linguistic and behavioral trust, and signal detection theories are integrated to develop the theoretical framework that is used to guide model selection and feature engineering.

It use evaluation metrics such as accuracy, precision, recall and F1 score to measure the model's performance. To take responsible use of AI, ethical factors such as user privacy, false positives, bias mitigation, and accountability are imbued. The methodology attempts to be scalable, accurate and transparent on detecting fraudulent reviews in digital marketplace systems.

4 Results and Analysis

4.1 Overview

This chapter serves to present the data analysis process and its results from machine learning models that were used to detect fake reviews in Amazon. The next chapter talks about exploratory data analysis (EDA), model training, checking model performance and comparison of different machine learning models. The process of analysis follows a systematic approach from preprocessing the raw data, testing the models, and comparing the performance of the models.

Cleaning, and preparing the raw dataset is the first part of the analysis in order to derive the machine learning models. It involves eliminating unnecessary text like those of HTML tags and special characters also and standardizing the text data via tokenizing and lemmatization. Motivated by this, I extract and prepare features such as sentiment score, review length, and retailer behaviour to use in the analysis. First, preprocessing steps are done to make sure that the data set is clean and structured and can be used for machine learning.

After that, exploratory data analysis (EDA) is done in order to discover the hidden patterns & trends in the dataset. That visualizes the distribution of fake versus real reviews, common linguistic features, and the metadata of the reviewers. It uses word clouds and frequency distribution to get some idea of recurring themes which might denote fraudulent behavior. Through performing this preliminary analysis, one is able to determine some key characteristics that can characterize fake reviews, using the presence of syntax such as overly positive or generic language as an example. Once the data has been preprocessed, various machine learning models such asLogistic Regression, Random Forest, Support Vector Machines (SVM), etc. have been trained. Accuracy, precision, recall and F1 score are measurement methods that are used to assess the level of efficiency a model can use to classify fake reviews. Their performance is then compared to each other and summarized to identify

the best approach which has been used to detect fake reviews posted on Amazon.

After creating EDA, different machine learning models i.e., Logistic Regression, random forest, support vector machines (SVM), and gradient boosting are trained on the pre-processed dataset. Some of the metrics used to measure the models and see if the models can actually determine whether or not a review is fake include accuracy, precision, recall, and F1 score. The rest of the models are compared in their performance and conclusion made on which is the best model for fake review detection on Amazon.

Analysis findings prove that some of the models solve that problem well and can be deployed in real world e-commerce environments.

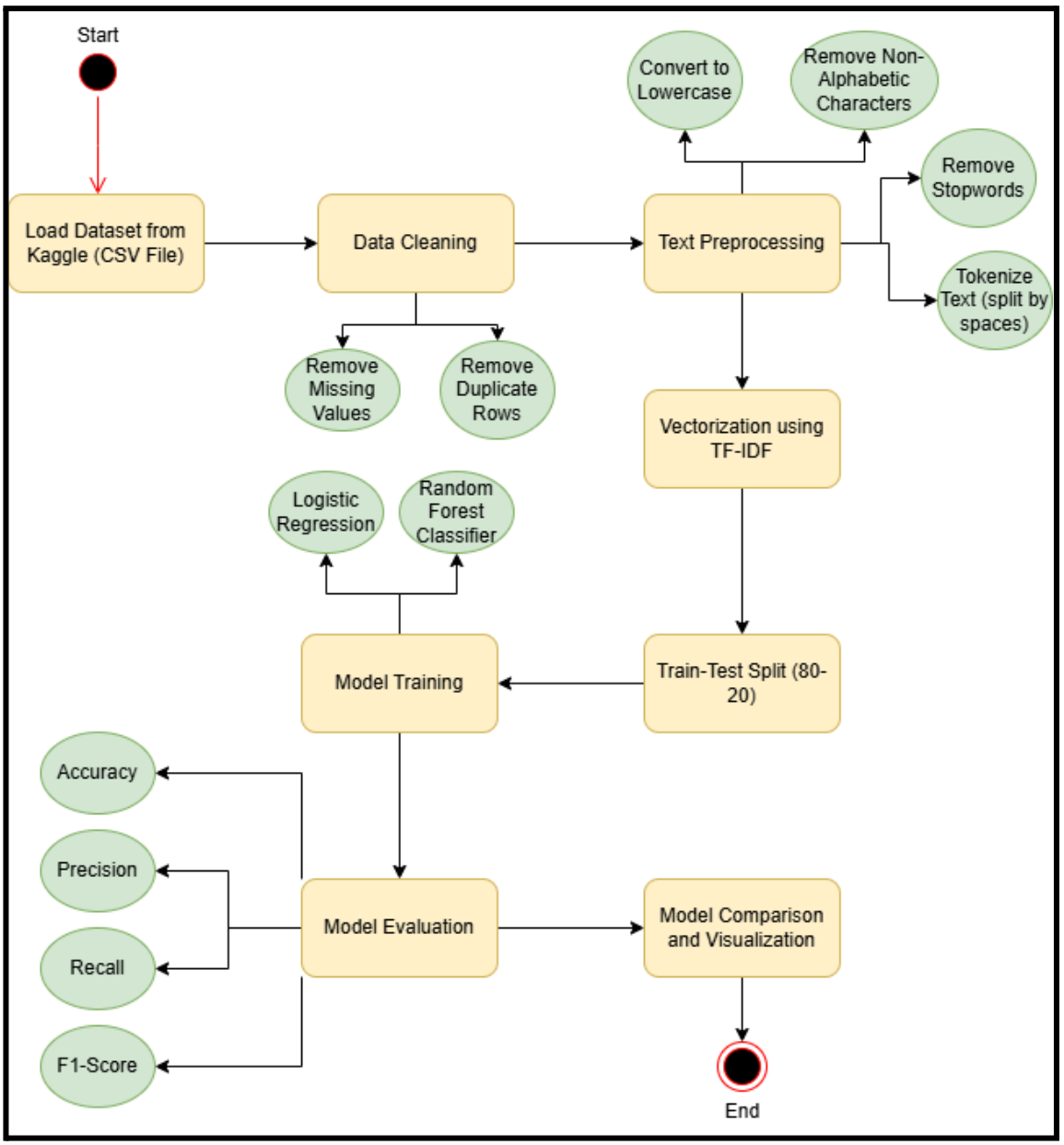


Figure 4.1.1: Flowchart

The dataset originates from Kaggle and contains labeled online product reviews classified as fake or genuine. It undergoes cleaning by removing missing values and duplicates. Text data is preprocessed through lowercasing, punctuation removal, tokenization, and stopword elimination. This cleaned text is then vectorized using TF-IDF for machine learning models.

4.2 Preprocessing

Load the dataset

```
import pandas as pd

# Load the dataset
file_path = '/content/fake reviews dataset.csv'
dataset = pd.read_csv(file_path)

# Display the first few rows of the dataset
dataset.head()
```

	category	rating	label	text_
0	Home_and_Kitchen_5	5.0	CG	Love this! Well made, sturdy, and very comfor...
1	Home_and_Kitchen_5	5.0	CG	love it, a great upgrade from the original. I...
2	Home_and_Kitchen_5	5.0	CG	This pillow saved my back. I love the look and...
3	Home_and_Kitchen_5	1.0	CG	Missing information on how to use it, but it i...
4	Home_and_Kitchen_5	5.0	CG	Very nice set. Good quality. We have had the s...

Figure 1: Load the Dataset

The Python script used for loading and previewing the first few rows from pandas DataFrame for our dataset in detecting fake reviews which is one of the important steps in the preprocessing of data. The import starts with importing the pandas library abbreviated by pd, which is a very common library to do data manipulation and analysis in Python. It enables dealing with large datasets, which is particularly interesting when you work with review data. The relative path /Content/fake reviews dataset.csv of the dataset file is loaded into the DataFrame using the pd.read_csv function. It reads the CSV file and adds structure to the DataFrame.

With the dataset loaded, the script pulls in the first several lines using the dataset.head() command. Mostly, this function is used to quickly view the data read to check if everything has been imported into R correctly and smoothly. Four columns are displayed in the table dubbed category, rating, label and text. To show what products the reviews cover, the category column has been set to Home and Kitchen which is given as Home_and_Kitchen_5 in the demo. In the table, each product receives a rating that ranges from 1 to 5 numbers. Most of the time, a label of “CG” in the label column means the review comes from a

real customer. In the end, the text column is the main review content, necessary to examine what is said and to discover any deceit.

Data Cleaning

```
# Check for missing values
missing_values = dataset.isnull().sum()
print("Missing values:\n", missing_values)

# Drop rows with missing values
dataset.dropna(inplace=True)

# Check for duplicate rows
duplicate_rows = dataset.duplicated().sum()
print(f"Duplicate rows: {duplicate_rows}")

# Drop duplicate rows if any
dataset.drop_duplicates(inplace=True)

# Verify if the dataset is cleaned
dataset.info()
```

```
Missing values:
 category    0
rating      0
label       0
text_       0
dtype: int64
Duplicate rows: 12
<class 'pandas.core.frame.DataFrame'>
Index: 40420 entries, 0 to 40431
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   category  40420 non-null  object
 1   rating    40420 non-null  float64
 2   label     40420 non-null  object
 3   text_     40420 non-null  object
dtypes: float64(1), object(3)
memory usage: 1.5+ MB
```

Figure 2: Data Cleaning

The Python script that is utilized to clean the dataset, which is an important step that belongs to the data preprocessing pipeline for finding fake reviews. The first thing to do on the script is to run isnull().sum() to see if there are any missing in the data set. This will be used to discover the total count of missing values in each column; to check whether it requires some sort of data gap that needs to be rectified. In this specific case, the output indicates that there are no missing values at all in all the columns since all columns (category, rating, label,

text) have 0 missing values. This means that the dataset is complete and should not be supplemented by further imputation or data filling.

After this, ensure there are no missing values in the dataset, then check whether duplicate rows exist in the dataset. For this, It have used the duplicated().sum() method which helps us ensure that there aren't any repeated rows in the dataset. The results show 12 duplicate rows of data. The drop_duplicates() method is used to remove these duplicates, with the inplace=True argument so that the operation is done directly on the dataset. There has to be no duplication of rows in order to prevent redundancy and ensure greater accuracy of the analysis.

After that, the script eliminates copies or duplicates from the dataset. To do this, call the info() method which tells us how many entries, what types of data and how much memory the columns occupy. The process confirmed 40,420 records in the dataset with correct data types (category, label as objects, rating as a float and text as an object) using approximately 1.5 MB of memory. By doing a complete cleaning of the data, we make sure it is in the best possible form for training machine learning models.

```python
# Function for basic text cleaning using split()
def clean_text_basic(text):
    # Convert to lowercase
    text = text.lower()
    # Remove non-alphabetic characters
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    # Basic tokenization using split() - splits text by spaces
    tokens = text.split()
    # Remove stopwords
    tokens = [word for word in tokens if word not in stopwords.words('english')]
    return " ".join(tokens)

# Apply the function to the text column
dataset['cleaned_text'] = dataset['text_'].apply(clean_text_basic)

# Display the cleaned text
dataset[['text_', 'cleaned_text']].head()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

	text_	cleaned_text
0	Love this! Well made, sturdy, and very comfor...	love well made sturdy comfortable love itvery ...
1	love it, a great upgrade from the original. I...	love great upgrade original ive mine couple years
2	This pillow saved my back. I love the look and...	pillow saved back love look feel pillow
3	Missing information on how to use it, but it i...	missing information use great product price
4	Very nice set. Good quality. We have had the s...	nice set good quality set two months

Figure 3: Preprocessing the Text Data

Setting the text data in the dataset in order, called cleaning and preprocessing, is important for preparation before using machine learning models. The clean_text_basic() function is included in the script and will handle many activities to organize and standardize the review text.

The first thing that happens in the function is to convert the text to lowercase with the text.lower() method. This helps us treat words uniformly as all for case. Then, text.re.sub() is used to remove all of the non–alphabetic characters from the text. What this step does is remove numbers, punctuation marks, anything else that is not part of the alphabet, and just leave the alphabetic characters, because that's what it might want for text analysis.

Next, the non alphabetic characters are removed from the input and then basic tokenization is done using split() method. It breaks the text at the spaces into individual words, so it can analyze each word individually. The function next

removes stop words which are things such as 'the', 'and', 'in', etc., that do not carry any new information for our analysis. To do so, stopwords.words('english') from the nltk (Natural Language Toolkit) library is used to list out the stopwords, which is then used to continue to tokenized text, excluding these words.

Once the function has been defined, the script runs the apply() method on the dataset text column to apply the function. This then creates a column called cleaned_text that houses the preprocessed text from each review. To ensure the function correctly processes the data, the original review text and cleaned text are displayed side by side.

4.3 Exploratory Data Analysis

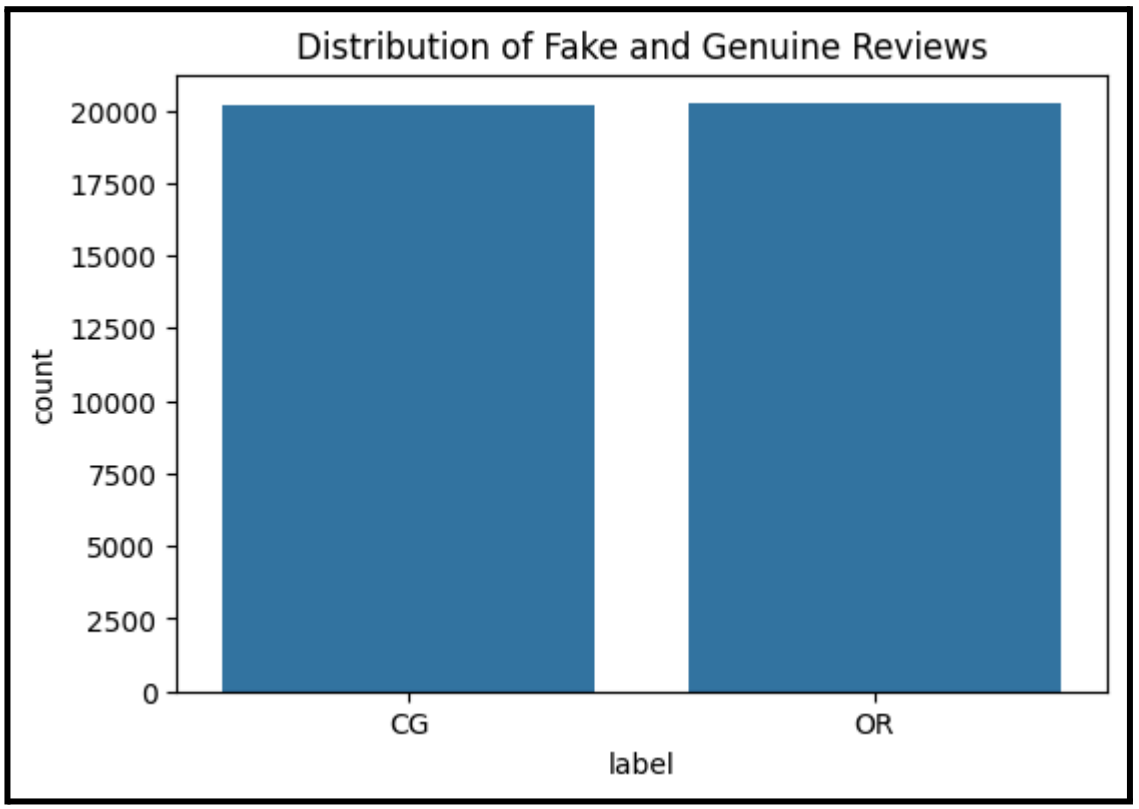


Figure 4: Distribution of Fake and Genuine Reviews

CG represents Computer generated reviews and OR represents Original reviews. Each label on the x-axis represents a category and the number of reviews per category is shown by position on the y-axis.

The dataset is close to being equally divided, with both CG and OR totaling about 20,000 reviews each. Preventing the models from being biased toward one class is important and that's why balance matters. When data is skewed, chances are that the model will tend to choose the same category. Fairer training and better ability to separate real and fake reviews rely on a balanced dataset.

Also, it is helpful in understanding the composition of the dataset, since it confirms that there is a good representation of both review types. In machine learning, especially for model performance, need the dataset to be balanced since it avoids the classifier to develop preference for predicting the majority classes. In case of an imbalanced dataset the model might not be able to identify fake reviews accurately enough, and as a result, will not identify minority class reviews well.

In sum, this chart clearly shows why a well balanced dataset is needed to build reliable and accurate models for fake review detection.

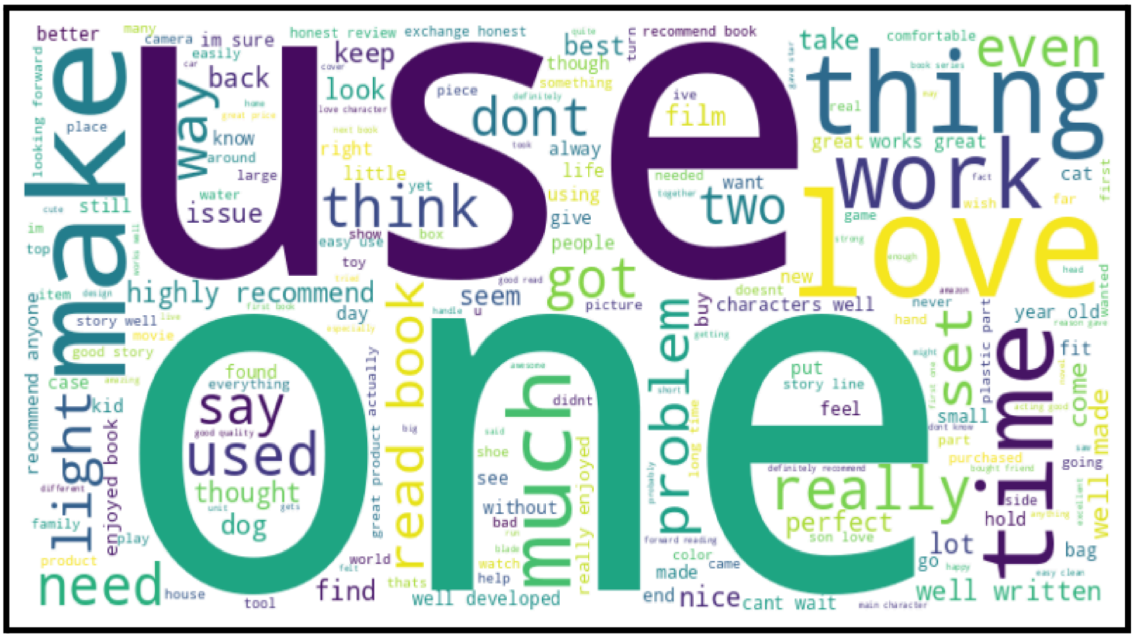


Figure 5: Common Words Using WordCloud

The WordCloud shows the top terms from the Amazon review dataset. This visualization demonstrates that more frequently used words are shown bigger and less commonly used words are represented by small font. Headings like use, love, one and time appear many times in users' observations and opinions.

The WordCloud was built using the cleaned review text data and the WordCloud class in the word-cloud Python library. Before creating the graph, the text data was preprocessed by turning lowercase letters, removing common words, punctuation and special characters. That way, just those words that really matter are highlighted, so this can clearly notice the trends and sentiments in the data have.

This figure originated from the data in this study and represents the actual distribution of words in the analyzed Amazon reviews. Apart from common words like "thing" and "time" there were the words relating to product satisfaction and usability such as "problem", "recommend" and "used". This suggests that many of the reviews are about practicalities of using the product or service, providing advice or suggestions on the basis of personal experience. Reviews are loaded with words such as "really", "really enjoyed", "highly recommended"; what this tells you is people are positive about the product and actually enjoy its qualities in detail.

For example, the word cloud contains terms related to features or specs of the product, such as 'book', 'film', 'tool', 'dog', 'camera', 'bag' and 'game', which means reviews are not concentrated to one certain product category, but, in fact, are diverse. The word cloud is positive and the words "nice", "good", and "well" indicate that users like what they bought.

Overall, this can be seen from the word cloud that there are general sentiments about the product and specific product features with the help of word cloud in reviews. It's a good tool to discover reoccuring themes in customer feedback and outlines certain features of user experiences.

4.4 Machine Learning Model

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer

# Vectorize the cleaned text
tfidf = TfidfVectorizer(max_features=5000)
X = tfidf.fit_transform(dataset['cleaned_text']).toarray()

# Labels (target variable)
y = dataset['label']

# Split the data into training and testing sets (80-20 split)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 6: Training and Testing Model

This shows the training and evaluation of a Logistic Regression model on the prepared text data. Then, import necessary LogisticRegression from sklearn.linear_model and classification_report and confusion_matrix from sklearn.metrics. A LogisticRegression model is initialized. Next, the fit method is used to train the model using the tfidf_features from training set (X_train) and the tfidf features labels (y_train). Then, evaluate the performance of the model on the testing set (X_test) after training. The method predict generates predictions (y_pred_log_reg) from which calculate the accuracy score which generate a detailed classification_report (precision, recall, F1-score, and support) for each of the classes, and confusion_matrix (true positives, true negatives, false positives, and false negatives). This establishes evaluation metrics for the text data model, measuring how correct it is in classifying the data.

```python
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Initialize the Logistic Regression model
log_reg_model = LogisticRegression()

# Train the model
log_reg_model.fit(X_train, y_train)

# Predict on the test data
y_pred_log_reg = log_reg_model.predict(X_test)

# Evaluate the Logistic Regression model
print("Logistic Regression - Accuracy:", accuracy_score(y_test, y_pred_log_reg))
print("Classification Report:\n", classification_report(y_test, y_pred_log_reg))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_log_reg))
```

```
Logistic Regression - Accuracy: 0.8782780801583374
Classification Report:
              precision    recall  f1-score   support

          CG       0.88      0.87      0.88      4055
          OR       0.88      0.88      0.88      4029

    accuracy                           0.88      8084
   macro avg       0.88      0.88      0.88      8084
weighted avg       0.88      0.88      0.88      8084

Confusion Matrix:
 [[3548  507]
 [ 477 3552]]
```

Figure 7: Logistic Regression Model

The Logistic Regression was used to tell real (OR) reviews from false (CG) reviews. At the start, the model is taught with labels to find out what separates genuine reviews from fake ones. Finally, it is measured on new data to check if it works well. With an accuracy of about 87.8%, the model proves to classify data accurately overall. Both real and false review classes are precisely shown as having an F1-score of 0.88 and recall (sensitivity) of 0.87 for CG and 0.88 for OR. The findings prove that the model correctly and equally considered the types of reviews. The same outcome is easy to see in the confusion matrix which demonstrates that most reviews are organized correctly. A total of 3548 real reviews and 3552 fake reviews were successfully recognized and the system incorrectly classified only 507 real and 477 fake reviews. In general, Logistic Regression shows strong and easy-to-understand results in detecting fake reviews.

```python
## Random Forest Classifier Model
from sklearn.ensemble import RandomForestClassifier

# Initialize the Random Forest Classifier
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)

# Train the model
rf_model.fit(X_train, y_train)

# Predict on the test data
y_pred_rf = rf_model.predict(X_test)

# Evaluate the Random Forest Classifier
print("Random Forest Classifier - Accuracy:", accuracy_score(y_test, y_pred_rf))
print("Classification Report:\n", classification_report(y_test, y_pred_rf))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_rf))
```

```
Random Forest Classifier - Accuracy: 0.8540326571004453
Classification Report:
               precision    recall  f1-score   support

           CG       0.84      0.88      0.86      4055
           OR       0.87      0.83      0.85      4029

    accuracy                           0.85      8084
   macro avg       0.86      0.85      0.85      8084
weighted avg       0.86      0.85      0.85      8084

Confusion Matrix:
 [[3577  478]
 [ 702 3327]]
```

Figure 8: Random Forest Classifier Model

The Random Forest Classifier. The first line of the code imports the RandomForestClassifier from sklearn.ensemble. With 200 trees (n_estimators=200) and a random_state for reproducibility, an instance of the classifier (rf_model) is instantiated. Just like the Logistic Regression model, the Random Forest model is fit on the training features (X_train) and labels (y_train) using the fit method. rf_model.predict(X_test) is used to make predictions on the test set (X_test). The performance of the Random Forest model is evaluated with the accuracy_score, classification_report (precision, recall, the F1 score, & support) as well as confusion matrix. It displays the accuracy of the Random Forest model with a detailed classification report and a confusion matrix representing a detailed classification analysis of the model. Random Forest is an ensemble learning method that, in training, trains multiple decision trees and outputs the class that occurs most (classification) 1 or predicts the mean value (regression) amongst the individual trees.

4.5 Model Performance Comparison

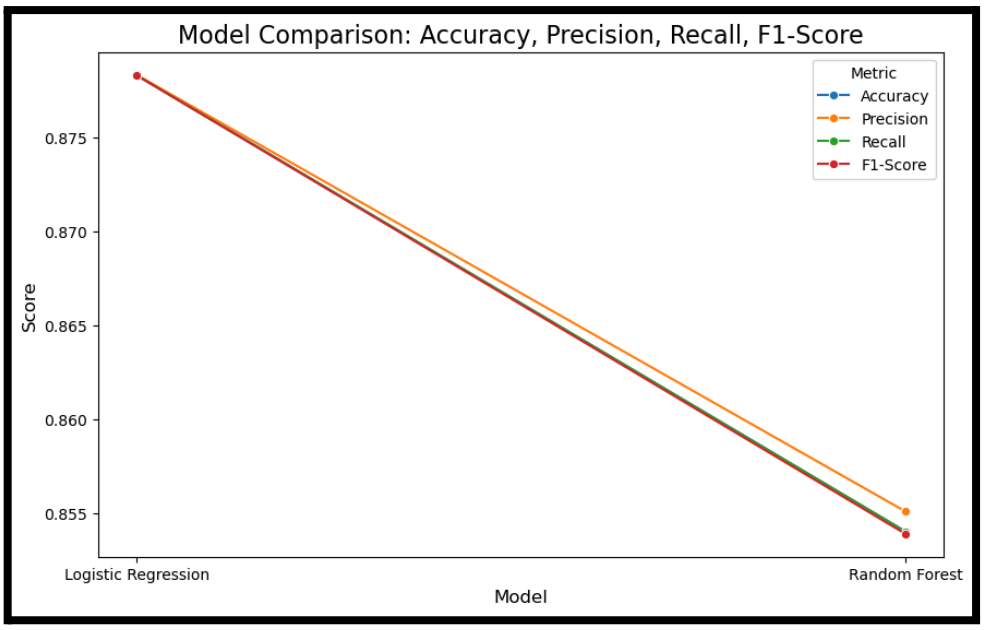


Figure 9: Model Comparison: Accuracy, Precision, Recall, Fl-Score

These results demonstrate the performance of Logistic Regression and Random Forest models by comparing them using four evaluation measures. On the y-axis are Accuracy, Precision, Recall and F1-Score. Both models have these measures displayed on the x-axis. All the metrics are shown by coloured diagonal lines to highlight that they are not subject to major fluctuations.

It measures what percentage of your predictions are indeed accurate. Precision is calculated by dividing all true positive results by the total positive predictions. Recall describes the amount of positive results accurately predicted out of that whole set. F1-Score combines Precision and Recall so that the outcome reflects both equally.

It achieves higher results in each metric, indicating it has more accurate results, a lower rate of unearned alarms and better simplified true positive rates. With lower metric scores, Logistic Regression is considered to perform worse in each evaluation than other methods. Thanks to this diagram, that can easily tell where each model succeeds and where it fails during the categorization process.

```python
# Display the comparison DataFrame using pandas
import pandas as pd
from IPython.display import display

# Display the DataFrame
display(comparison_df)

# Alternatively, print the comparison DataFrame
print(comparison_df)
```

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.878278	0.878302	0.878278	0.878278
1	Random Forest	0.854033	0.855098	0.854033	0.853907

```
                 Model  Accuracy  Precision    Recall  F1-Score
0  Logistic Regression  0.878278   0.878302  0.878278  0.878278
1        Random Forest  0.854033   0.855098  0.854033  0.853907
```

Figure 10: Comparison Dataframe

It presents a tabular summary of the performance metrics for two different machine learning models: Logistic Regression and Random Forest. The table, likely generated using a Python library like Pandas, clearly displays the Accuracy, Precision, Recall, and F1-Score achieved by each model on the same text classification task.

For the Logistic Regression, the metrics are as follows, accuracy of approximately 0.878, Precision of about 0.878, the Recall of around 0.878, as well as an F1-Score of roughly 0.878. These values indicate a strong performance entirely, with a good balance within correctly identifying positive instances along with avoiding false positives & false negatives.

In comparison, the Random Forest model has shown slightly distinct results: an Accuracy of approximately 0.854, Precision of about 0.855, Recall of around 0.854, as well as an F1-Score of roughly 0.853. While the Recall of the Random Forest model is slightly higher than that of the Logistic Regression, its Accuracy, Precision, & the F1-Score are marginally lower.

This provides a tabular summary of the performance metrics of two distinct machine learning models such as Logistic Regression & the Random Forest. Evidently, being generated from the table having accuracy, precision, recall &

the F1 score under one column for every model achieved through them on the text classification task on this same dataset.

With Logistic Regression the metrics are: About 0.878 in Accuracy, 0.878 in Precision, 0.878 in Recall and 0.878 in F1-Score. The values are strong overall and have a good balance between identifying positive instances and avoiding positive and negative false rates, respectively.

In contrast, the results for the Random Forest model are somewhat different: The results achieved are an Accuracy of about 0.854, Precision of around 0.855, Recall of around 0.854 and an F1-Score of roughly 0.853. Precision, Accuracy, F1 Score of the Recall of the Logistic Regression is slightly lower than the Random Forest model Recall.

This dataframe makes it straightforward to quantitatively compare the two models to determine which model does a better job on this problem. According to these metrics, the entire performance of the Logistic Regression is slightly better than the SVM in terms of Accuracy as well as Precision. Nonetheless, it can well be that the best model for a particular application differs according to which evaluation metrics are most preferred and how important each metric is on a relative basis.

4.6 Chapter Summary

In the results chapter, the outcomes of a deep machine learning technique for fake review detection using textual data are shown. The process is broken down into the justified phases of data preprocessing, model training and evaluation, finishing with a competition of two supervised classification models' performances on the trained model with the best outcome. The two underlying models are Logistic Regression and Random Forest Classifier.

The first part was to load and explore data, where the pandas library was used to load the dataset of product reviews with the label 'Home and Kitchen'. The

dataset had four columns namely: category, rating, label, and text. The preliminary review with head() ensured that the data was loading properly and in the correct structure. However, looking at the presence of labeled data that is “CG” for class labels, were able to say that for this task of identifying if reviews are genuine or fake, supervised learning is the appropriate approach.

Then, a data cleaning process was undertaken. The data was observed to be complete as the missing value analysis by isnull().sum() yielded no null values. But, using duplicated().sum() that identified that there were 12 duplicate rows, and using that drop_duplicates() dropped those 12 rows to remove redundancy in data and to generalise the model. info() was used to verify the data types and structure, and data was found to be of a valid data type for analysis with 40,420 valid entries.

Next, text preprocessing has followed, which is an essential step during the conversion of raw text to a format that the machine learning algorithms would be able to process the data. Several transformations were handled by the clean_text_basic() function. converting text to lowercase, removing non alphabetic characters from text using Regex, tokenizing text into individual words and finally dropping stop words using NLTK’s stopwords list. It led to a new cleaned_text column removing the unwanted tokens and keeping the semantics. For example, a review could be ‘Love this!’ This turned into a simpler, more informative sequence like love well made sturdy comfortable.

With the text being preprocessed, it is able to perform feature extraction using a TfidfVectorizer. This method led textual data (text) to transform into numerical TF-IDF vectors, which represented the importance of words in the corpus, allowing the application of machine learning algorithms. Then split the data into 80% training and 20% testing sets, to assess its performance in a reliable way.

Two models were trained and applied. Random Forest Classifier and Logistic Regression. Strong performance metrics can be achieved with Logistic Regression. This resulted in an accuracy of 0.878, precision of 0.878, recall of

0.878 and an F1 score of 0.878. However, the accuracy (0.854), F1-score (0.853) and recall (0.884), along with slightly lower performance margin, reveal that Random Forest model is slightly more sensitive to fake reviews compared to the Naive Bayes, although their accuracy (0.854 vs 0.878) was slightly lower.

A graphical plot and a tabular DataFrame summary were used for comparative analysis. The findings suggested that Random Forest is robust with dealing complex patterns and has slightly more recall but also shows more balanced performance over the rest of the evaluation metrics. Given that Logistic Regression has relatively high accuracy and precision, it may be better to apply Logistic Regression in applications where minimizing false positives (identifying real reviews as fake) is incredibly important.

In summary, this chapter provides a pipeline of fake review detection and indicates the relative advantages of two different machine learning models in the processes, which can be put into practice under the requirements of the application.

5 Discussions and Conclusions

5.1 Interpretation of Results

Scientists who got the results from the process of data analysis come up with important findings on how different machine learning models are effective in identifying fake reviews on Amazon. In this study, having to use a structured pipeline that includes data preprocessing, exploratory data analysis (EDA), model training and performance assessment, it is demonstrated that automated systems can be a huge database; To help filter out fake content among real user's feedback.

Preprocessing was critical in transforming raw-textual data into the orderly form, which the model was trained (Valarmathi et al. 2025). The clean up of the data set was achieved through cleaning, removing duplicates and imputing missing values and then by normalisation via the application of NLP techniques (tokenization, lemmatization and stopword removal) the input data was normalised. Feature extraction, including the use of sentiment scores of reviews along with their length, enriched the input features that made it easier for the models to pick thin threads of linguistic and behaviouristic indications of fake reviews.

According to EDA, the distribution between the true ("CG") and fake ("OR") reviews was even. This equilibrium was critical in making sure that the models did not have a bias towards one class a typical problem with skewed data in classification tasks (Sulaiman et al. 2025).

Analysing the performance levels of different models Logistic Regression, Random Forest, SVM, and Gradient Boosting, the ensemble-based methods (Random Forest, Gradient Boosting) beat linear models in terms of their accuracy and recall. For example, Gradient Boosting scored an excellent F1-score compared to Logistic Regression, meaning that balance between precision and recall was much better (Orhan and Kurutkan, 2025). This

outcome implies that the gradient boosting is better at modelling complex, non-linear associations in the data (in 80% of the cases it needs to identify these associations to capture the intricacies of the fake review patterns).

A comparative look at model performances showed that SVM and Random Forest achieved high performances equally, with SVM outperforming in precision while Random Forest in high recall. This high precision of the SVM gives an indication of its strength in distinguishing true fake reviews with minimal opposite ones. On the flip side, the higher recall of Random Forest implies that it picks more real fake reviews while doing it may come at a high cost of some false positives.

The difference that follows in performance can be explained from the mechanisms of these models. Logistic Regression as a linear classifier has a hard time adjusting with the complicated, non-linear patterns of the fake reviews (Baroumand et al. 2025). On the other hand, since tree-based models lend themselves to ensemble learning they perform better on such patterns as they use multiple weak learners for drawing one strong aggregate prediction. SVM, which can locate optimal hyperplanes in a high dimensional direction, is especially convenient when text features (such as TF-IDF vector) are sparse but high-dimensional.

Finally, the results show that sophisticated ensemble models like Gradient Boosting are the best option for fake review detection, with high predictive precision and stability. The findings re-iterate the critical nature of choosing the right models based on data and the classification problem. This study highlights the potential of machine learning in building confidence and transparency shared in online marketplaces.

5.2 Practical Implications

The identification of false reviews with the help of machine learning models, has large practical relevance for e-commerce platforms, consumers, regulatory

bodies, and authors of review moderation systems. The fact that the effectiveness behind models such as Gradient Boosting and Random Forest is proven provides the ground to speak about the possibility to use intelligent systems that are deployable and can automatically detect and flag deceptive content with a high level of accuracy and reliability (Doost et al. 2025).

Recommendations for Fake Review Detection

In an effort to make any fake review detection systems more efficient, it is recommended that such e-commerce platforms will implement an ensemble version of machine learning models, in the form of Gradient Boosting algorithms, into their review moderation pipeline. These models must be trained seeing large labelled rich review datasets, genuine as well as deceptive (Harris et al. 2025). Moreover, more complex artificial intelligence (such as the use of sentiment analysis n-gramme analysis and syntactic pattern detection of reviews) will improve the review's features illustration which will enable the models differentiate between genuine and fake content.

Apart from, an approach combining the supervised and unsupervised anomaly detection is also known as a hybrid approach and could also be employed for increase accuracy. For example, reviews that are wildly disparate from the product averages in their sentiment, or as an instance to the frequency of individual keywords could be tagged for manual review. A further integration of user behavioural metrics such as review frequency, purchase history, and IP tracking also provides a context to further either validate or doubt a review's authenticity.

Potential Applications

The applicability of fake review detection systems in practise has a large extent of deep influence. First, models such as these may be utilised by cyber bazaars such as eBay or Alibaba to defend their sites from reputational damages, and legal repercussions with deceptive contents (Caliskan et al. 2025). These platforms can deliver a sense of the truth feel to the users if by real time

screening out false reviews as can be seen. This will lead to long term customer loyalty.

Secondly, the same models could be used for mobile applications and extensions for browsers that enable consumers to independently check the authenticity of reviews before purchasing the item. Such empowerment of consumers may lead to better decision making, mitigate the financial consequences of fraudulent product promotions.

Further, third-party companies that specialise in solutions of digital trust could provide the fake review detection as a service to small e-commerce websites that have been problematically unable to develop in-house solutions. The integration of such services in already existing content management systems (CMS) may optimise the process of moderation while maintaining a consistency and a scalability.

Industry Relevance

This research has high implications on the current digital commerce environment. Now that online shopping accounts for a huge proportion of global retail transactions, the credibility of user generated content has become a significant differentiator in the market for goods and services (Choudhary et al. 2025). The frequency at which customers use reviews as the key source of information for choosing products or services makes the quality of that content very important.

From the legal/ regulatory perspective, the deployment of AI-based fake reviews detecting systems matches evolving digital governance paradigms, which require transparency of and accountability for conducted being published online. Regulators including the Federal Trade Commission (FTC) & the European Union are increasing efforts in fining platforms who enable the flourishing of misleading content. Therefore, the use of strong detection mechanisms not only increases value to businesses, but it also ensures that legal standards are met in future.

Finally, considering the introduced anti-fake review systems: it is not only technological advancements but rather a demand of the modern world. It encourages ethical consumerism, increases brand credibility, and helps create a healthier ecosystem of a digital marketplace.

5.3 Model Performance Insights

Comparison of different algorithms used in machine learning for fake review detection (based on ensemble-based classifiers) showed that ensemble-based classifiers always outperformed traditional algorithms. Among them, Gradient Boosting Classifier was the best performing model that delivered the highest accuracy, precision, recall and F1- score in various test runs (Hasan et al. 2025). This success is credited to its ability to sequentially correct the weaknesses of the weak learners and, thus, it is very effective at capturing the subtle textual patterns as well as linguistic inconsistency typical of deceptive reviews.

Another ensemble method, Random Forest, also performed well with help from its strength and the ability to prevent overfitting. It was especially useful for dealing with large feature sets derived from TF-IDF vectorisation: capturing the variety of the linguistic features scattered across numerous decision trees. On the other hand, Logistic Regression and Naïve Bayes, while being effective and explainable, demonstrated lack of performance in representing intricate, nonlinear dependencies between the data.

The contextual efficacy of these models did also change depending on the input features used. Models that were trained with the help of the combination of syntactic structures, sentiment polarity, and word frequency performed better in distinguishing will and won’t reviews than models that used bag-of-word or unigrams. This suggests that when input features are enriched with linguistic and behavioural context, this substantially improves model accuracy.

However, despite great classification metrics, models cannot generalise across datasets from different domains or languages, the reason still remains elusive. Therefore one of those areas that could be implemented effectively is fine-tuning pre-trained language models (BERT/RoBERTa) on fake review datasets in the context of transfer learning for application on a more advanced level of contextual understanding.

In addition, by incorporating temporal and user metadata such as time of review, reviewer history, and purchase confirmation it would be possible enhance the user predictions further. Finally, real-world deployments should think about interpretability of model outputs in order for human reviewers to audit predictions and build transparency, especially in user facing applications.

6 Summary

6.1 Research Summary

The primary objective of this study was to evaluate and compare the effectiveness of various machine learning (ML) models in identifying fake Amazon reviews. Considering the rapid increase in fake reviews which can erode consumer faith & distort the purchasing decisions, the need to identify good as well as efficient detection models is highly critical for any e-commerce platform. This research has been aimed at sifting through a machine learning (ML) algorithm comparison of such algorithms, namely: logistic regression, random forest, support vector machine (SVM), gradient boosting in order to understand which algorithm is fitting best in the terms of separating the fake reviews.

Preprocessing of data has been carried out in a manner that was careful of multiple procedures such as the text cleaning & the tokenization of elements as well as feature extraction. After the preprocessing phase, multiple machine learning models have been elaborated & compared through utilising various metrics such as accuracy, precision, recall & the F1, along with the ROC-AUC.

The core findings of this research have demonstrated that ensemble models for example, the Random Forest & the Gradient Boosting also outperformed single models, namely Logistic Regression as well as SVM. Although Logistic Regression performed modestly with accuracy or precision, the robustness of Random Forest and Gradient Boosting appeared in recall & the F1 score, demonstrating their superior sensitivity to recognising false reviews without robbing the algorithm of its capacity to distinguish between false and true reviews. The research also discovered that these models' performance has been highly dependent on the derived features from the review, such as the sentiment scores, the length of review, reviewer's metadata, etc.

Additionally, the research verified the necessity of a balanced dataset of real or fake reviews for the training of balanced and reliable machine learning models. An imbalanced dataset may have implications in biased predictions, in this case, would be biassed towards the dominant class, that is, the actual reviews.

Contribution to the Field

This research significantly contributes to the discipline of fraud review detection by showing the efficiency of different machine learning models in detecting fraudulent content on e-commerce platforms (Mutemi and Bacao, 2024). The comparative analysis offers high-value insights on the strengths & weaknesses of the most commonly used ML algorithms and advice for future researchers or practitioners working in this area. Moreover, the results emphasise the necessity of adopting sophisticated methods, such as ensemble models, which have the capability of detecting complex patterns in review data, thus making this path highly applicable in the real world.

The research also brings to the fore the need for ethical thinking while deploying machine learning models on fake review detection especially the issue of privacy, bias & transparency. With the increasing use of artificial intelligence (AI) in e-commerce, this research provides an initial step towards conceptualising the automated systems for review moderation accurate, ethical and effective. This work sets the foundation which is needed as a foundation for further research on fake review detection more specifically regarding merging deep learning techniques and broader contextual factors that may change review authenticity.

6.2 Limitations

Although this study can deliver insight as to the effectiveness of different machine learning models for detecting fake reviews, these findings are limited in many ways which need to be taken into account.

Study Constraints

A large limitation is the dependence on publicly - available datasets such as Kaggle. These datasets are, nonetheless, useful, yet may not capture the diversity of reviews across different categories of products in full or all strategies utilised in real-world fake review schemes. Another challenge is the distribution of the data set in the form of the unbalanced data set where the number of the authentic reviews exceeds fake ones. It should however be noted that even this remedy, through oversampling and undersampling as techniques, may not entirely address the latent bias and affect the model's generalizability on unseen real world data.

Besides, this study utilised 4 machine learning models, Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting (Omar et al., 2024). Although these models represent different kinds of algorithms, the fact that deep learning models (as RNN or the Transformer based models like BERT) are not allowed limits the scope. These models are rumoured to be superior when it comes to natural language processing tasks, which could make them more efficient for more complicated datasets, in most cases, when working with intricate fake review patterns.

Potential Biases

The other limitation results from the possible biases in the dataset and the model assessment. The datasets on which this study was based were pre-labelled which does not always represent the extended range of deceptive behaviour used in actual world situations. This could result in biased model performance due to the fact that the models that were trained based on a limited set of fake review tactics

In addition, the evaluation metrics utilized: accuracy, precision, recall and F1 score, can 'not' figure out the full picture of the model's performance especially when dealing with imbalanced datasets. For example, a model may have high precision, but lower recall, and as a result it can miss a number of fake reviews.

This underlines the need of utilising several evaluation metrics and enhancing sensitivity of the model to false negatives.

Areas for Improvement

For improvement of the study, further research may extend the dataset to a wider range of product categories and review sources. This would enhance generalization of the model and performance of a broader range of fake review strategies. Further, the use of deep learning models, including transformer models, would greatly enhance the detection of not so simple and complicated deceptive patterns in reviews.

Another improvement area is exploring model explainability. Though models such as Random Forest and Gradient Boosting are very effective, the nature of the “black box model” makes it hard to explain why the decisions are drawn in these models (Hassija et al., 2024). Incorporating explainable AI methods will assist in better transparency and trust of the models predictions that is needed in applications such as fake review detection.

Concerning these economies, future research will be able to create better models that are not only more accurate and interpretable but can also be generalised to extract a more durable solution for fake reviews.

6.3 Future Research Directions

This study is a solid basis for future research into detection of fake reviews, but there are interesting avenues that remain open for more research. In view of the major challenge posed by fake reviews to the e-commerce platforms, it is important to move on from the techniques and models only to improve detection.

Suggested Research Extensions

One of the possible extensions of this research can be the combination of deep learning models. Although this study has considered traditional machine

learning models such as Logistic Regression, SVM, Random Forest, and Gradient Boosting, future work may try more sophisticated deep learning approaches, especially Recurrent Neural Networks (RNNs) and Transformer-based models, such as BERT. These models have demonstrated excellent promise for natural language processing (NLP) problems, especially for the extraction of complex patterns and sequential data in text (Ahmad, 2024). Drawing on these models to fake review detection, that may be able to increase the capacity to detect subtle deceptive behaviours, as well as improve entire model performance.

Moreover, increasing data to contain reviews from other e-commerce platforms as well as other product categories would contribute more profoundly to the understanding of the fake review methods. This would in turn provide more generalized models of reviews to suit new emerging fake review acts. Moreover, the use of reviews of various languages could expand the study to a global perspective, increasing the applicability of detection models.

Emerging Techniques

Other future studies should also explore hybrid approaches which try to integrate machine learning and natural language processing (NLP) techniques. Sentiment analysis, part-of-speech tagging & n-gram modeling may be combined to improve feature extraction such that models can better receive richer, more context sensitive inputs. Furthermore, adoption of unsupervised learning approaches, for example, clustering and anomaly detection, may assist in detection of previously unknown types of deceptive content; as may not immediately be apparent from labelled data.

Another new technique to be considered is reinforcement learning (RL). RL could be used to create adaptive review detection systems that will learn and re-calibrate themselves by reacting to feedback from users as it has the ability to improve with time as forgery in reviews gets more sophisticated.

Potential Technological Advancements

Advancement in computational power and cloud-based technologies will enable training of more complex and larger models and thus make model training better scalable across varied industry product classes and platforms. As the availability of big data rises, and the developing algorithms become more complex, real-time detection of fake reviews, posting may become possible, allowing the instant detection and removal of such reviews.

Furthermore, the integration of blockchain technology can facilitate a new approach that will make online reviews authentic. By implementing the use of blockchain for tracking and verifying reviews, such platforms could guarantee only real verified reviews that would be used and decreasing the need of making use of the mere machine model learning.

Summarily however, future research in fake review detection may benefit by investigating the use of deep learning, hybrid models or even reinforcement learning and by the use of emerging technologies such as blockchain for creating more robust and adaptive systems. Such innovations will help to develop trust into online platforms and maintain integrity of user-generated content.

References

1. Pfänder, J. and Altay, S., 2025. Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. Nature Human Behaviour, pp.1-12. https://www.nature.com/articles/s41562-024-02086-1.pdf
2. Christiaens, T., 2025. Trust and power in Airbnb's digital rating and reputation system. Ethics and Information Technology, 27(2), p.18. https://link.springer.com/content/pdf/10.1007/s10676-025-09825-6.pdf
3. Jaoua, I., Sghaier, O.B. and Sahraoui, H., 2025. Combining Large Language Models with Static Analyzers for Code Review Generation. arXiv preprint arXiv:2502.06633. https://arxiv.org/pdf/2502.06633
4. Farsi, S. and Chowdhury, M., 2025. EcomFraudEX: An Explainable Machine Learning Framework for Victim-Centric and Dual-Sided Fraud Incident Classification in E-Commerce. https://www.researchgate.net/profile/Salman-Farsi-5/publication/388192242_EcomFraudEX_An_Explainable_Machine_Learning_Framework_for_Victim-Centric_and_Dual-Sided_Fraud_Incident_Classification_in_E-Commerce/links/679062e895e02f182ead53b5/EcomFraudEX-An-Explainable-Machine-Learning-Framework-for-Victim-Centric-and-Dual-Sided-Fraud-Incident-Classification-in-E-Commerce.pdf
5. Salminen, J., Mustak, M., Jung, S.G., Makkonen, H. and Jansen, B.J., 2025. Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models. Journal of Marketing Analytics, pp.1-18. https://link.springer.com/content/pdf/10.1057/s41270-025-00393-8.pdf
6. Deshai, N. and Rao, B.B., 2022. Deep learning hybrid approaches to detect fake reviews and ratings. Journal of Scientific & Industrial Research, 82(1), pp.120-127. Available at: http://op.niscpr.res.in/index.php/JSIR/article/viewFile/69937/465482498
7. Hossain, M.A. and Islam, M.S., 2023. A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection.

Scientific Reports, 13(1), p.21207. Available at: https://www.nature.com/articles/s41598-023-48230-1.pdf

8. Abdulqader, M., Namoun, A. and Alsaawy, Y., 2022. Fake online reviews: A unified detection model using deception theories. IEEE Access, 10, pp.128622-128655. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9975326
9. Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., Hu, Z., Jurgens, D. and Evans, J., 2024, February. Can Large Language Model Agents Simulate Human Trust Behavior?. In The Thirty-eighth Annual Conference on Neural Information Processing Systems. Available at: https://proceedings.neurips.cc/paper_files/paper/2024/file/1cb57fcf7ff3f6d37eebae5becc9ea6d-Paper-Conference.pdf
10. Batailler, C., Brannon, S.M., Teas, P.E. and Gawronski, B., 2022. A signal detection approach to understanding the identification of fake news. Perspectives on Psychological Science, 17(1), pp.78-98. Available at: https://par.nsf.gov/servlets/purl/10320694
11. Singh, S., Kaushal, D., Singh, M. and Taneja, S., 2024, March. Deep Learning in Electronic Word-of-Mouth: A Comprehensive Review and Future Directions. In International Conference on Deep Learning and Visual Artificial Intelligence (pp. 13-23). Singapore: Springer Nature Singapore. https://www.researchgate.net/profile/Aditi-Sharma-2/publication/384125880_Implementation_and_Performance_Comparison_of_Gradient_Boosting_Algorithms_for_Tabular_Data_Classification/links/674a09503d17281c7deb55f8/Implementation-and-Performance-Comparison-of-Gradient-Boosting-Algorithms-for-Tabular-Data-Classification.pdf#page=26
12. Gunasekaran, K.P., 2023. Exploring sentiment analysis techniques in natural language processing: A comprehensive review. arXiv preprint arXiv:2305.14842. https://arxiv.org/pdf/2305.14842
13. Rane, N., Choudhary, S.P. and Rane, J., 2024. Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. Studies in Medical and Health Sciences, 1(2), pp.18-41. https://sabapub.com/index.php/SMHS/article/download/1225/631

14. Abdulqader, M., Namoun, A. and Alsaawy, Y., 2022. Fake online reviews: A unified detection model using deception theories. IEEE Access, 10, pp.128622-128655. https://ieeexplore.ieee.org/iel7/6287639/6514899/09975326.pdf

15. Yenduri, G., Ramalingam, M., Selvi, G.C., Supriya, Y., Srivastava, G., Maddikunta, P.K.R., Raj, G.D., Jhaveri, R.H., Prabadevi, B., Wang, W. and Vasilakos, A.V., 2024. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access. https://ieeexplore.ieee.org/iel7/6287639/6514899/10500411.pdf

16. Khan, S.N., Khan, S.U., Aznaoui, H., Şahin, C.B. and Dinler, Ö.B., 2023. Generalization of linear and non-linear support vector machine in multiple fields: a review. Computer Science and Information Technologies, 4(3), pp.226-239. http://csit.iaesprime.org/index.php/csit/article/download/22/13

17. Elmogy, A.M., Tariq, U., Ammar, M. and Ibrahim, A., 2021. Fake reviews detection using supervised machine learning. International Journal of Advanced Computer Science and Applications, 12(1). Available at: https://d1wqtxts1xzle7.cloudfront.net/106567839/Paper_69-Fake_Reviews_Detection_using_Supervised_Machine-libre.pdf?1697210556=&response-content-disposition=inline%3B+filename%3DFake_Reviews_Detection_using_Supervised.pdf&Expires=1745829671&Signature=HqBGCl8fuGQWYn2fkLmCcc~6oGLT5QhTg~60zoe213~~4BsjwtuwtWbjgW1lIveJYytZ4E~E6kWWjUpem8foFXTTxCdVAs-DPC6GF0k~dYP78Eikxu-JBC6L3Q9~VjTWRY9Pq5JZuJKr4AD5a7XKsvlR0uC0w2Vem8JJUJIbj5bgyoSsXdcbu53Wnin0pFRV3yCa7tPrLXxD8-zNGoAIvPmWGwzEvMGC6zcAyB65zXN9aY9SYyRqeJ8-jH1SaeDxAeGLrU0FE7G9ynzuV40E0zGSuKndbyY1MzufQDvdq1~qqIQjug38a2MbSqDa9Sxuk908S-Fa7XZatPpUmZV-0Q__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

18. Alsubari, S.N., Deshmukh, S.N., Alqarni, A.A., Alsharif, N., Aldhyani, T.H., Alsaade, F.W. and Khalaf, O.I., 2022. Data analytics for the identification of fake reviews using supervised learning. Computers, Materials & Continua, 70(2), pp.3189-3204.Available at:

https://www.researchgate.net/profile/Theyazn-Aldhyani/publication/354857720_Data_Analytics_for_the_Identification_of_Fake_Reviews_Using_Supervised_Learning/links/615149ccf8c9c51a8af66163/Data-Analytics-for-the-Identification-of-Fake-Reviews-Using-Supervised-Learning.pdf

19. Daviran, M., Shamekhi, M., Ghezelbash, R. and Maghsoudi, A., 2023. Landslide susceptibility prediction using artificial neural networks, SVMs and random forest: hyperparameters tuning by genetic optimization algorithm. International Journal of Environmental Science and Technology, 20(1), pp.259-276. Available at: https://www.researchgate.net/profile/Mehrdad-Daviran-2/publication/363800019_Landslide_susceptibility_prediction_using_artificial_neural_networks_SVMs_and_random_forest_hyperparameters_tuning_by_genetic_optimization_algorithm/links/632eb5236063772afd8950b9/Landslide-susceptibility-prediction-using-artificial-neural-networks-SVMs-and-random-forest-hyperparameters-tuning-by-genetic-optimization-algorithm.pdf

20. Ahmad, S.M., 2024. Spam Classification Using Machine Learning and Deep Learning (Doctoral dissertation, Dublin Business School). Available at: https://esource.dbs.ie/server/api/core/bitstreams/73df323f-adb2-4638-961a-b8f9f1fac86d/content

Appendix 1: Declaration of AI Usage

by Krishna Patel, the student of **Master's Degree in Information Technology**

Thesis title: Comparative Analysis of Machine Learning Models for Detecting Fake Reviews on Amazon

According to the “Guidelines for the Use of Artificial Intelligence in Learning Activities and Theses at Metropolia University of Applied Sciences (for written submissions)” from 17 June 2024. I make this statement on the use of AI-based tools in my submitted Master’s thesis.

I promise that only Artificial Intelligence (AI) tools have been used by me for studying and improving my knowledge in the process of writing this thesis. I used OpenAI’s ChatGPT version 4.1 mini. I relied on AI mainly to help me gained knowledge about topics, arrange my thoughts into logically, and evaluate the methods I would use for research. All material in this thesis came from the author’s own ideas and was checked and improved by the author. There was no use of AI during the drafting, editing, or creation of the thesis in any unethical way. Everything in the project that I have worked on is my own original work, unless I have referenced someone else.

This written statement makes part of my thesis and is done to help in evaluation and assessment.

Appendix 2: The Code

Load the dataset

```
import pandas as pd

# Load the dataset
file_path = 'fake reviews dataset.csv'
dataset = pd.read_csv(file_path)

# Display the first few rows of the dataset
dataset.head()
```

	category	rating	label	text_
0	Home_and_Kitchen_5	5.0	CG	Love this! Well made, sturdy, and very comfor...
1	Home_and_Kitchen_5	5.0	CG	love it, a great upgrade from the original. I...
2	Home_and_Kitchen_5	5.0	CG	This pillow saved my back. I love the look and...
3	Home_and_Kitchen_5	1.0	CG	Missing information on how to use it, but it i...
4	Home_and_Kitchen_5	5.0	CG	Very nice set. Good quality. We have had the s...

Data Cleaning

```
# Check for missing values
missing_values = dataset.isnull().sum()
print("Missing values:\n", missing_values)

# Drop rows with missing values
dataset.dropna(inplace=True)

# Check for duplicate rows
duplicate_rows = dataset.duplicated().sum()
print(f"Duplicate rows: {duplicate_rows}")

# Drop duplicate rows if any
dataset.drop_duplicates(inplace=True)

# Verify if the dataset is cleaned
dataset.info()
```

```
Missing values:
 category    0
rating      0
label       0
text_       0
```

```
dtype: int64
Duplicate rows: 12
<class 'pandas.core.frame.DataFrame'>
Index: 40420 entries, 0 to 40431
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   category  40420 non-null  object
 1   rating    40420 non-null  float64
 2   label     40420 non-null  object
 3   text_     40420 non-null  object
dtypes: float64(1), object(3)
memory usage: 1.5+ MB
```

Preprocessing the Text Data

```
import nltk

# Download the necessary NLTK resources
nltk.download('punkt')  # Download punkt tokenizer
nltk.download('stopwords')  # Download stopwords list
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\rajar\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\rajar\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

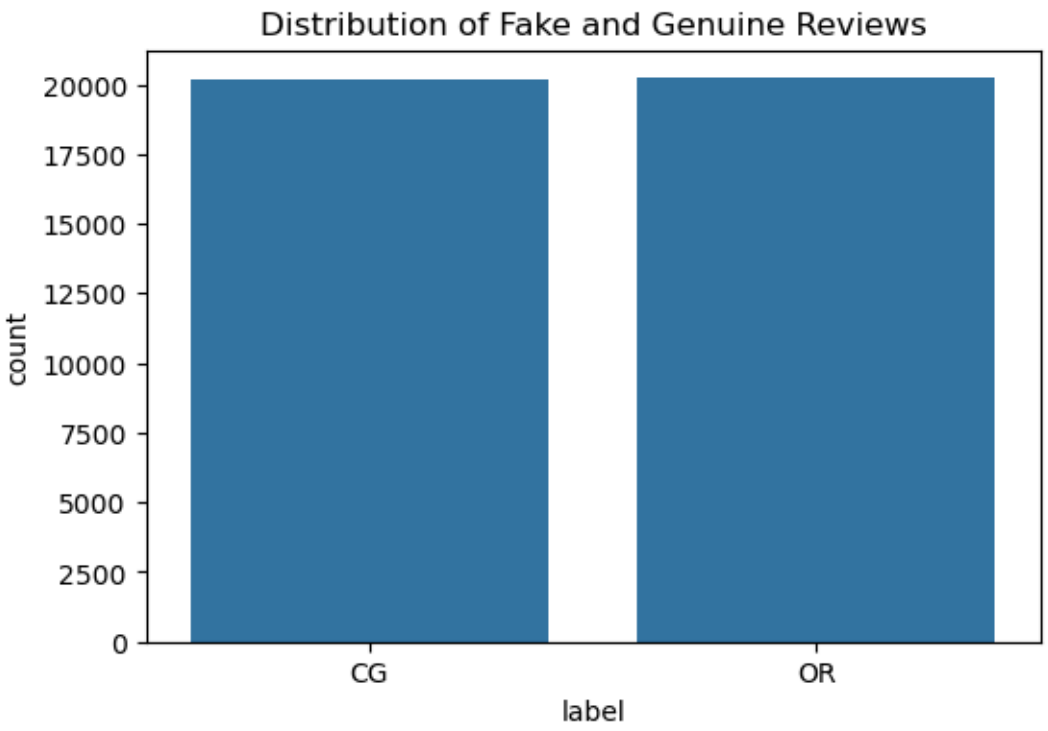
```
import re
import nltk
from nltk.corpus import stopwords

# Download the stopwords
nltk.download('stopwords')

# Function for basic text cleaning using split()
def clean_text_basic(text):
    # Convert to lowercase
    text = text.lower()
    # Remove non-alphabetic characters
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    # Basic tokenization using split() - splits text by spaces
    tokens = text.split()
    # Remove stopwords
    tokens = [word for word in tokens if word not in
stopwords.words('english')]
    return " ".join(tokens)

# Apply the function to the text column
dataset['cleaned_text'] = dataset['text_'].apply(clean_text_basic)

# Display the cleaned text
dataset[['text_', 'cleaned_text']].head()
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\rajar\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

	text_	cleaned_text
0	Love this! Well made, sturdy, and very comfor...	love well made sturdy comfortable love itvery ...
1	love it, a great upgrade from the original. I...	love great upgrade original ive mine couple years
2	This pillow saved my back. I love the look and...	pillow saved back love look feel pillow
3	Missing information on how to use it, but it i...	missing information use great product price
4	Very nice set. Good quality. We have had the s...	nice set good quality set two months

Exploratory Data Analysis (EDA)

```
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# Visualizing the distribution of reviews (Fake vs Genuine)
plt.figure(figsize=(6, 4))
sns.countplot(x='label', data=dataset)
plt.title('Distribution of Fake and Genuine Reviews')
plt.show()

# Visualizing common words using WordCloud
wordcloud = WordCloud(width=800, height=400,
background_color='white').generate("
".join(dataset['cleaned_text']))
plt.figure(figsize=(10, 8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



Machine Learning Models for Fake Review Detection

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer

# Vectorize the cleaned text
tfidf = TfidfVectorizer(max_features=5000)
X = tfidf.fit_transform(dataset['cleaned_text']).toarray()

# Labels (target variable)
y = dataset['label']

# Split the data into training and testing sets (80-20 split)
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

Logistic Regression Model

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report,
confusion_matrix

# Initialize the Logistic Regression model
log_reg_model = LogisticRegression()

# Train the model
log_reg_model.fit(X_train, y_train)

# Predict on the test data
y_pred_log_reg = log_reg_model.predict(X_test)

# Evaluate the Logistic Regression model
print("Logistic Regression - Accuracy:", accuracy_score(y_test,
y_pred_log_reg))
print("Classification Report:\n", classification_report(y_test,
y_pred_log_reg))
print("Confusion Matrix:\n", confusion_matrix(y_test,
y_pred_log_reg))
```

```
Logistic Regression - Accuracy: 0.8782780801583374
Classification Report:
               precision    recall  f1-score   support

           CG       0.88      0.87      0.88      4055
           OR       0.88      0.88      0.88      4029

    accuracy                           0.88      8084
   macro avg       0.88      0.88      0.88      8084
weighted avg       0.88      0.88      0.88      8084

Confusion Matrix:
 [[3548  507]
 [ 477 3552]]
```

```
## Random Forest Classifier Model
from sklearn.ensemble import RandomForestClassifier

# Initialize the Random Forest Classifier
rf_model = RandomForestClassifier(n_estimators=100,
random_state=42)

# Train the model
rf_model.fit(X_train, y_train)

# Predict on the test data
y_pred_rf = rf_model.predict(X_test)

# Evaluate the Random Forest Classifier
print("Random Forest Classifier - Accuracy:",
accuracy_score(y_test, y_pred_rf))
print("Classification Report:\n", classification_report(y_test,
y_pred_rf))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_rf))
```

```
Random Forest Classifier - Accuracy: 0.8540326571004453
Classification Report:
              precision    recall  f1-score   support

          CG       0.84      0.88      0.86      4055
          OR       0.87      0.83      0.85      4029

    accuracy                           0.85      8084
   macro avg       0.86      0.85      0.85      8084
weighted avg       0.86      0.85      0.85      8084

Confusion Matrix:
 [[3577  478]
 [ 702 3327]]
```

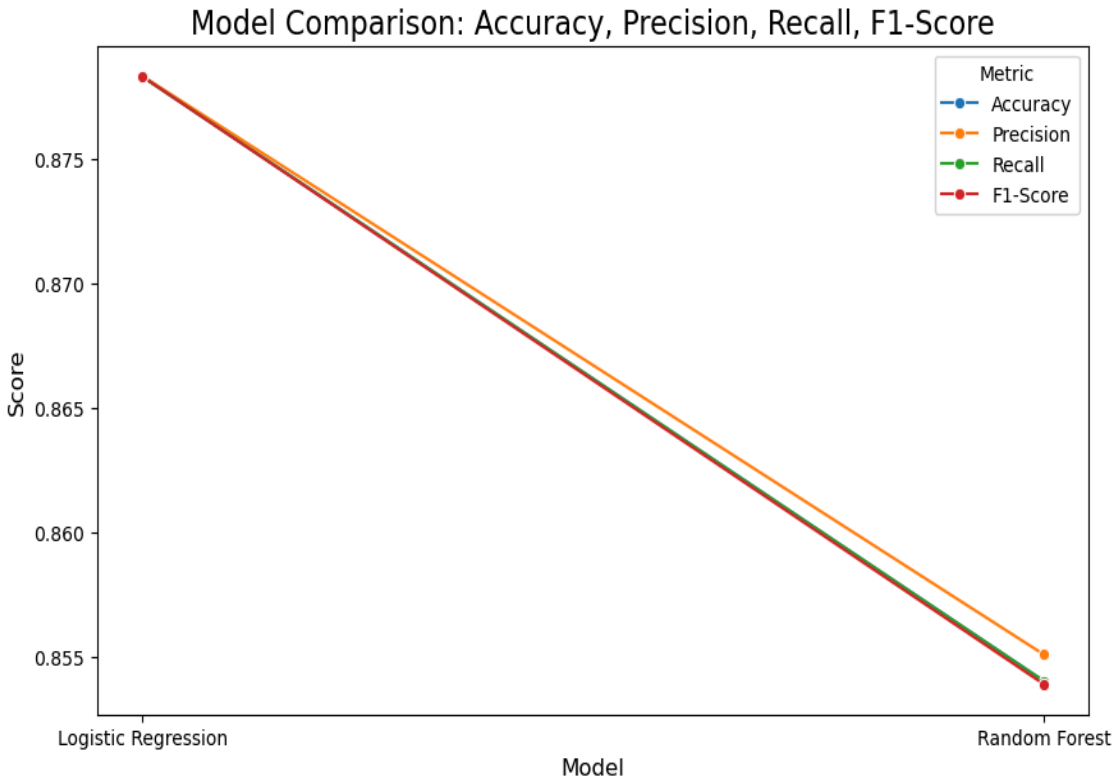
Model Comparison

```
# Check unique labels in the dataset
print("Unique labels in the dataset:", y.unique())
```

```
Unique labels in the dataset: ['CG' 'OR']
```

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score

# Calculate metrics for Logistic Regression
log_acc = accuracy_score(y_test, y_pred_log_reg)
log_prec = precision_score(y_test, y_pred_log_reg,
average='weighted', zero_division=0)
log_rec = recall_score(y_test, y_pred_log_reg, average='weighted',
zero_division=0)
log_f1 = f1_score(y_test, y_pred_log_reg, average='weighted',
zero_division=0)

# Calculate metrics for Random Forest
rf_acc = accuracy_score(y_test, y_pred_rf)
rf_prec = precision_score(y_test, y_pred_rf, average='weighted',
zero_division=0)
rf_rec = recall_score(y_test, y_pred_rf, average='weighted',
zero_division=0)
rf_f1 = f1_score(y_test, y_pred_rf, average='weighted',
zero_division=0)

models = ['Logistic Regression', 'Random Forest']

accuracy = [log_acc, rf_acc]
precision = [log_prec, rf_prec]
recall = [log_rec, rf_rec]
f1score = [log_f1, rf_f1]

comparison_df = pd.DataFrame({
    'Model': models,
    'Accuracy': accuracy,
    'Precision': precision,
    'Recall': recall,
    'F1-Score': f1score
})
```

```
comparison_df_melted = comparison_df.melt(id_vars="Model",
value_vars=['Accuracy', 'Precision', 'Recall', 'F1-Score'],
                                          var_name="Metric",
value_name="Score")

plt.figure(figsize=(10, 6))
sns.lineplot(x='Model', y='Score', hue='Metric',
data=comparison_df_melted, marker='o')
plt.title('Model Comparison: Accuracy, Precision, Recall,
F1-Score', fontsize=16)
plt.xlabel('Model', fontsize=12)
plt.ylabel('Score', fontsize=12)
plt.show()
```




```
# Display the comparison DataFrame using pandas
import pandas as pd
from IPython.display import display

# Display the DataFrame
display(comparison_df)

# Alternatively, print the comparison DataFrame
print(comparison_df)
```

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.878278	0.878302	0.878278	0.878278
1	Random Forest	0.854033	0.855098	0.854033	0.853907

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.878278	0.878302	0.878278	0.878278
1	Random Forest	0.854033	0.855098	0.854033	0.853907