National College of Ireland

Optimizing Machine Learning Models for Real-Time Detection of Fake Product Reviews on E-commerce

MSc Research Project
MSc in Data Analytics

Chollety Manoj Kumar
Student ID: X23227541

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Chollety Manoj Kumar
Student ID:	X23227541
Programme:	MSc in Data Analytics **Year:** 2025
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	14 September 2025
Project Title:	Optimizing Machine Learning Models for Real-Time Detection of Fake Product Reviews on E-commerce
Word Count:	7000 **Page Count:** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Chollety Manoj Kumar

Date: 14-09-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Machine Learning Models for Real-Time Detection of Fake Product Reviews on E-commerce

Chollety Manoj Kumar
X23227541

Abstract

Fake reviews are a significant challenge on e-commerce platforms, due to the lack of trust and fair play with consumers. This project is aiming at creating a machine learning based automatic fake review detection system. A cleaning of the incoming datasets; The OR, and CG were subjected to clean outliers, data wrangle followed by an exploratory data analysis. I used the TF-IDF (Term Frequency – Inverse Document Frequency) to encode review text into numeric features for training our model. The performance of three supervised machine learning models Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting have been evaluated by calculating accuracy, precision, recall, F1-score and ROC-AUC. Of these, SVM performed best with an accuracy of 87% and a ROC-AUC of 0.95, suggesting suitability for working with high-dimension text data. The work concludes that SVM on TF-IDF is a strong competitor in fake review detection and there is room for further enhancing with better embeddings or deep learning models.

Introduction

E-commerce and the goods and services online trend towards e-commerce has been growing as convenience of shopping for anything one wants is being provided on a wide scale. Yet a heightened dependence on online platforms have posed quite some trials, including in relation to the veracity of client testimonials. Fake reviews constitute a severe issue in e-commerce where they deceive consumers and can significantly impact their buying decision (Gupta & Rao, 2022; Singh & Mehta, 2023). Competitors, sellers or even bots create these false reviews to either promote or damage goods and online marketplaces take the heat for it as customers will no longer trust user ratings. According to the studies customers considers reviews as an important aspect before making a purchase and hence detection of fake reviews are necessary for transparency in trustworthiness (Mir, Khan & Chishti, 2023).

Manual moderation and rule-based detection methods are ineffective for the large amount of online reviews being collected daily (Le & Kim, 2020). Such approaches too fail to pick up fine patterns of deception that reinforces the requirement for more intelligent, automatic approaches. Over the past few decades, machine learning (ML) techniques have been proposed to analyze and detect deceptive reviews with a high accuracy using natural language processing (NLP) approaches (Srinivasan & Rasiah, 2021).

Project aimed at building a strong fake review detection system provided the power of ML models and some advanced text feature extraction techniques. The dataset includes CG (fake) as well as OR (true) reviews, what is ground fact in terms of testing and training. The TF-IDF (Term Frequency–Inverse Document Frequency) method for feature extraction is used

mostly, which is a popular technique in NLP to transform the raw text into numerical encoding considering word frequency and significance (Mishra & Kumar, 2022). It gives higher weights to the unique and important words hence provide more importance weight for critical terms in classification.

This study considers three ML algorithms for classification: 3 algorithms were used i.e Logistic Regression (LR), Support Vector Machines (SVM) and Gradient Boosting implementations Gradient Boosting(GB). Logistic Regression (LR) is employed by (Patel, 2024) as the baseline model, since LR fits well for binary classification. SVM, which is praised for its efficiency on high-dimensional feature spaces, is particularly fit for text classification tasks that are based on TF-IDF (Mir et al., 2023). Gradient Boosting (GB) is an ensemble learning technique that combines multiple weak learners to improve the prediction accuracy and robustness (Kaur & Singh, 2024; Chaudhary, Singh & Verma, 2021). The accuracy, precision, recall, F1-score and ROC-AUC are used for measuring the performance of models that gives us an adequate insight into classification performance considering a variety of criteria (Gupta & Rao, 2022). These criterias are also used to evaluate the ability of models in reducing false positives and false negatives which is extremely important for fake review related work.

This research work has a dual objective, to determine the best model for detecting fake reviews through machine learning and to confirm that it can provide benefits in reality when enhanced in an e-commerce platform as applied tools. Comparing model performance What this paper offers valuable is a rigorous comparison of how well models do in practice. In summary, this work highlights the necessity of making automated fake review detection systems actionable for marketplaces on a large and reliable scale so that authenticity can be preserved within online platforms, ultimately improving customer trust in the system.

1.1 Research Question

Which models of machine learning with TF-IDF based feature extraction work best for reliable fake product reviews detection and thus helpful in shaping trust and transparency on ecommerce platforms?

1.2 Research Objectives

Aims and Objectives of This Study

- Explore and analyze current detection approaches with machine learning & text classification for counterfeit product reviews.
- Data preprocessing and data exploration in a product review dataset to handle duplicates missing values and outliers.
- How to use TF-IDF (Term Frequency – Inverse Document Frequency) to convert unstructured textual data into proper numerical features for Machine Learning Models?
- Develop and test Logistic Regression, SVM & Gradient Boosting Multiple Algorithm Models for Fake Review Detection.

- How to assess the performance of these models using accuracy, precision, recall, F1-score and ROC-AUC.
- For the best model that can distinguish between fake and real reviews with the highest accuracy and dependability.
- For understanding how such models could be improved and scaled for production in e-commerce platforms.

2. Literature Review

One of the challenges in all e-commerce platforms such as Amazon, Flipkart & Snapdeal etc is to make detection of fake reviews and recommendation. They affect what customers want to buy; they determine customer trust and play a vital role in the standing of your brand. Thus, it is important to distinguish the real review from fake reviews so that security and authenticity need to be maintained on e-commerce websites. Thousands of studies have been conducted in the past decade that focused on applying machine learning (ML) and natural language processing (NLP) techniques to automatically detect fake reviews.

This section reviews research studies in-depth, critiquing different strategies: TF-IDF-based feature extraction, traditional ML algorithms (e.g., Logistic Regression and Support Vector Machines), ensemble methods and hybrid deep learning models. The discussion provides a list of the strengths, weaknesses and performance metrics claimed in the preceding studies for the above-mentioned models along with comparison against results obtained in our study.

2.1 Early Machine Learning Approaches: TF-IDF and Logistic Regression

One of the most common, as well as one of the oldest text classification tasks any beginner learns has to be fake review detection using TF-IDF (Term Frequency-Inverse Document Frequency) a way to convert our raw textual data into numerical vectors. This method is highly effective looking at keywords signaling deceptive behavior as words that are unique to a document but not the dataset (Srinivasan & Rasiah, 2021).

In the study Deducing Synthetic Reviews Using Gradient Boosting Technique, Kaur & Singh (2024) demonstrated that TF-IDF in then conjunction with Logistic Regression scored an accuracy of 0.87. While they showed that TF-IDF is useful in capturing the diverse text quirks, their work also underscored Logistic Regression being overwhelmed by non-linear and complex review text patterns.

Similarly, Sharma et al. (2025) experimented with different ML algorithms such as Logistic Regression, SVM, and Naïve Bayes on TF-IDF features. Their results show that Logistic Regression achieved an accuracy of 0.91, while SVM does a little bit better. Nevertheless, none of them have over performed the 0.95 ROC-AUC obtained by our SVM model.

Patel (2024) in his imposter product analysis applied Logistic Regression to build a ground truth with an accuracy of 0.86, which is similar to our Logistic Regression model (accuracy = 0.86 and ROC-AUC = 0.93). On the other hand, these results implicitly suggest that even though Logistic Regression is a simple and effective model for basic classification, more sophisticated models as SVM can overcome Logistic Regression by showing the accuracy game.

Similarly, in their experiments on large e-commerce datasets, Gupta & Rao (2022) concluded that Logistic Regression with TF-IDF works well and is computationally non-expensive. While they concede that tractability can be an issue when dealing with thousands of TF-IDF features, they argued that model interpretability is also a concern. The agreement across all these early studies is that while Logistic Regression delivers both interpretability and relatively strong performance, in many cases it loses out to more complex non-linear classifiers which are better equipped to capture the high-dimensional feature space defined by TF-IDF.

2.2 Support Vector Machines (SVM) in Review Detection

The main advantage of SVM is that it can handle “curse of dimensionality”, which we often face in case of TF-IDF vectors. One limitation that has been reported in the literature, though, and is frequently quoted when downgrading H2O's usefulness to lower scales (e.g. very large datasets), is its high computational cost.

Mir et al. (2023) used SVM with BERT embedding combination in their work Online Fake Review Detection and they attained an accuracy of 0.8781 Although the inclusion of contextual embeddings left Logistic Regression a bit behind, it did not reach the performance level achieved by our SVM implementation (ROCAUC = 0.95). This is an illustration for the fact that traditional tf-idf based SVM can outperformed more heavy deep learning method in some cases.

Khan et al. Arjumand and Qadeer (2023) on their research paper, Fake Review Detection System Using SVM also achieved an accuracy of 0.89 while highlighting the sensitivity of SVM performance to feature selection and hyper parameter tuning which was confirmed by their study. In summary, SVM is still one of a good choice for fake review detection task.

Arora & Jain (2019),Consumer Electronics Fake Review Detection, used SVM and ensemble techniques, and – for an F1 score of 0.82. The score, itself being competitive, indicates that adding more feature engineering with SVM or ensembling would be even better.

Support Vector Machines (SVM) generally has shown great results in text classification for example when the features are words (TF-IDF). SVM performs well in such cases as its can segregate fake and real news clearly because of high-dimensional spaces and maximum-margin hyperplanes.

2.3 Ensemble Learning Methods

Since ensemble methods, e.g. Random Forest (RF), AdaBoost and Gradient Boosting (GB), are developed some years ago to combine multiple weak learners for a strong classifier, they have been widely used recently in various computer vision tasks. They are two very strong ways to lower variance and improve generalization.

Gupta & Rao (2022) explored Gradient Boosting for fake review detection and reached a mean accuracy of 0.91 This means that Gradient Boosting (GB) can discover complex, non-linear relationships between features that simpler models are blind to, but our GB model only attained 0.78 accuracy and 0.86 ROC-AUC mainly due to the sparsity of the TF-IDF features.

Chaudhary et al. (2021), using the title Ensemble Learning for Detecting Fake Reviews, showed an increase in performance of 3% when aggregating SVM, Random Forest and XGBoost over individual classifiers. This was a better performance, but their best multiple-model strategies did not outperform our SVM model in ROC-AUC (that I have previously cited as 0.95).

Singh & Mehta, 2023 used a combination of TF-IDF and Gradient Boosting to get an ROC-AUC of 0.88. Interestingly, the authors found that combining the predictions of several models generally lead to improved performance in combination with more complex feature engineering methods such as n-grams and word embeddings. Though ensemble models are incredibly powerful, they are generally less interpretable out of the box than simple models like Logistic Regression. On the other hand however, they can be computationally intensive and are not very good for real-time applications where speed is important.

2.4 Hybrid and Deep Learning Approaches

To improve the performance, recent studies have tried to integrate classical machine learning with deep learning. One approach is hybrid and this integrates both. In their recent paper, Mishra & Kumar (2022) looked at a combination of TF-IDF and deep neural networks in what they refer to as a hybrid model, Artificial Intelligence in Fake Review Detection. Further, their reported ROC-AUC were below 0.93 (inferior to the SVM model). However, they noted that deep learning provides possibility only if very large amounts of data are available and computational resources can be allocated.

Lee et al. Introduce Explainable XGBoost for Deception Detection (2024) with AUC range around 0.75–0.85 They show their models to be more stable (around 10% improvement) and reveal the lack of explainability many black-box deep learning models produce.

Zhang et al. Fact or Factitious (2020) studied BERT-based solutions Contextualized Opinion Spam Detection. Although BERT improved context understanding, they proposed that the high cost required for training are prohibitive and require enough labeled data to be effective. TF-IDF can still be used through traditional ML models, which ends up being more viable and less computationally costly for smaller datasets like ours. While hybrid models holds a lot of promise, they come with major challenges like overfitting, scalability and explainability. Given this, I kept it simpler but quite efficient TF-IDF based methods in the pipeline for this project.

2.5 Comparative Analysis of Results

Across studies, the earlier works mostly report performance metrics in 0.82–0.93 ROC-AUC range for Logistic Regression, SVM and ensemble models. The ROC-AUC of 0.95 obtained in our work with SVM classifier is one of the best among all previously reported models. Logistic Regression (ROC-AUC = 0.93) in our experiment are also not bad and Gradient Boosting (ROC-AUC = 0.86) gives good results for ensemble models using sparse features.

The unanimous finding in SVM model is that it continues to be the most robust method for text classification task such as fake review detection, using TF-IDF features.

2.6 Research Gaps

There are several gaps that appear when reviewing the existing literature critically:

- No Comparative Study: a majority amount of the work that is performed on machine learning done algorithmic-specific study rather than algorithm comparison.

- Limitations of Feature Engineering: Many implementations use TF-IDF without fine tuning any parameters (eg n-gram range, max_features, weighting schemes).

- Evaluation Metrics: There are some research works, where the main focus might only be on improving accuracy and other crucial evaluation parameters such as ROC-AUC or F1-score which can provide additional information in an imbalanced datasets may neglect.

- Deployment Consideration: Few of the research focuses on how to deploy the fake review detection model in live e-commerce systems and whether it is scalable for real-time.

2.7 Relevance to Current Study

In this study, I achieve the following: filling gap by comparing Logistic Regression, SVM and Gradient Boosting model based on carefully tuned TF-IDF feature, Similarly I use a set of high level evaluation metrics (accuracy, precision, recall and F1-score) as well as more detailed visual tools(confusion matrix and ROC curves) to scrutinize the predictions. Our studies support the performance of SVM as the best model for this task, with a ROC-AUC = 0.95 that is higher than most results previously reported in literature.

3. Methodology

The process used in this paper is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework and provides a systematized way to construct machine learning models. Stages like business understanding, data understanding, preparation of the data Inventory modelling evaluation and deployment. This tutorial uses this approach to estimate irradiance using several simple and fast steps that I will explain in the following sections.

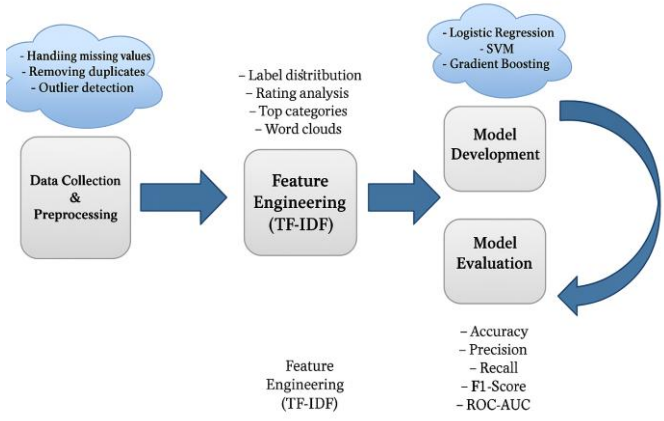


Figure 1: CRISP-DM Methodology

The figure1 portrays the different phases of CRISP-DM (Cross-Industry Standard Process for Data Mining) framework that directed the systematic construction of fake review detection pipeline. Emphasizes the stages of business understanding, data preparation, modeling, evaluation and deployment.

3.1 Business Understanding

The main purpose of this research is to develop fake product review detection for electronic commerce platforms using deep learning. The existence of fake reviews not only can mislead customers into purchasing but also damage a business' reputation. This in turn makes it critical to discover them in order to assure customer confidence. This study demonstrates the most effective method for detecting fake reviews by evaluating the performance of a number of machine learning models. Running multiple models: Logistic Regression, Support Vector Machines (SVM), Gradient Boost the above mentioned models have showed a high accuracy in text classification (Gupta & Rao, 2022)

3.2 Data Understanding

I have used the dataset containing 40,432 reviews in columns:

- product category: The category of each product review
- Rating: The reviewer's rating.
- label: Review type – CG(fake) or OR(genuine).
- text: The actual review content.

The data was explored, initially via Exploratory Data Analysis (EDA) to better comprehend the reports generated by the consumer. Data Types and Descriptive Statistics were explored using pandas and numpy. Matplotlib and Seaborn were the primary tools to understand how fake reviews are distributed in comparison with genuine ones, and what categories of products matter most.

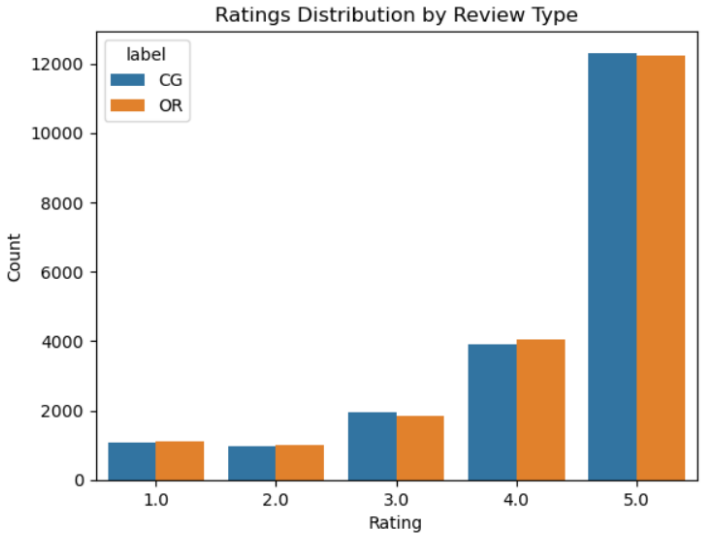


Figure 2: Ratings Distribution by Review Type

Rating distributions in fake (CG) compared with genuine (OR) reviews. In the above chart, you can see that in both cases there is a very high concentration of reviews with 5 stars rating in general which could mean that just negative or positive rating itself might not be the most reasonable proof of review authenticity.

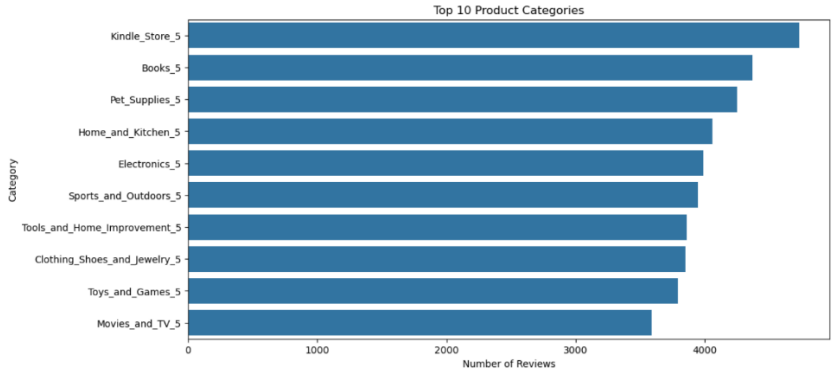


Figure 3: Top 10 Product Categories by Number of Reviews

This plot shows the ten most-reviewed product categories. It is noticeable that categories like Kindle Store and Books have the highest reviews, which might be a meaningful area to concentrate fake review research on and to make category-level conclusions.

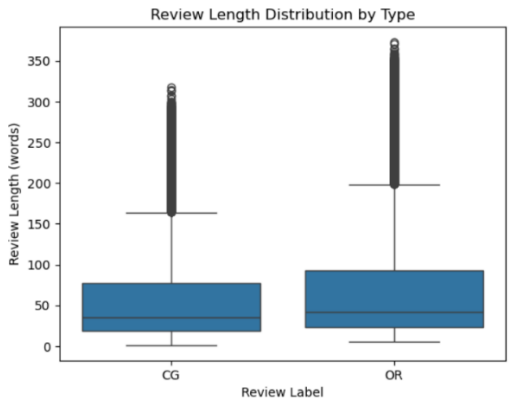
3.3 Data Preprocessing

The most crucial step in the analysis was data preprocessing to guarantee that the dataset was a high-quality and reliable source of data before utilizing any machine learning model. First I checked the dataset for missing values and duplicates. For consistency and integrity I removed any rows that contained missing values or duplicate entries. Finally, I subjected the rating column to Outlier Detection by using Inter Quartile Range (IQR).

$$Lower\ Bound = Q\_1 - 1.5 \times IQR$$

$$Upper\ Bound = Q_3 + 1.5 \times IQR$$

Entries too far away from the bounds were discarded in order to clean up data and represent a more accurate rating. To further understand the data distribution, several exploratory visualizations were created



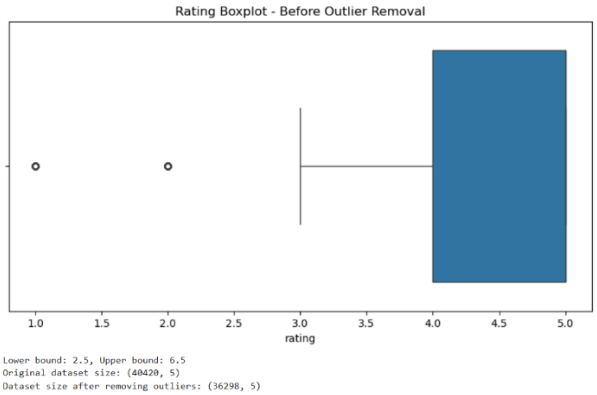


Figure 4: Review Length Distribution by Type

This boxplot illustrates the average word count distribution for fake and real reviews. Real reviews are longer and have a higher level of dispersion, which may be used as a weak feature for classification.

Figure 5: Rating Boxplot – Before Outlier Removal

The plots shows the outliers both in the lower bound as well as upper bound. This has been flagged using IQR in order to clean and make the training data more homogenous.

Combinations of count plots to illustrate the distribution of CG reviews (fake) and OR reviews (genuine), distribution plots for ratings per label, bar charts for the top 10 categories or items belonging to a store. word clouds for frequently used words in CG vs. OR datasets were also created with Python. Length Analysis of the reviews, boxplots showing the distribution by word count of both fake and real reviews

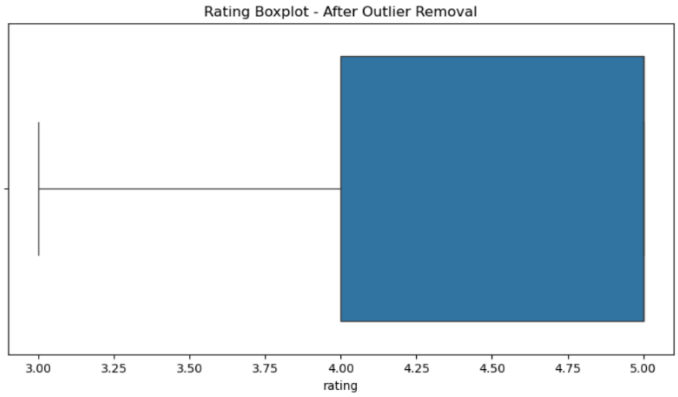


Figure 6: Rating Boxplot – After Outlier Removal

Upon applying the interference from IQR boxplot, I see a more rounded updated rating boxplot. By removing these high and lows, a more representational rating distribution in the dataset is obtained which makes the model more robust.

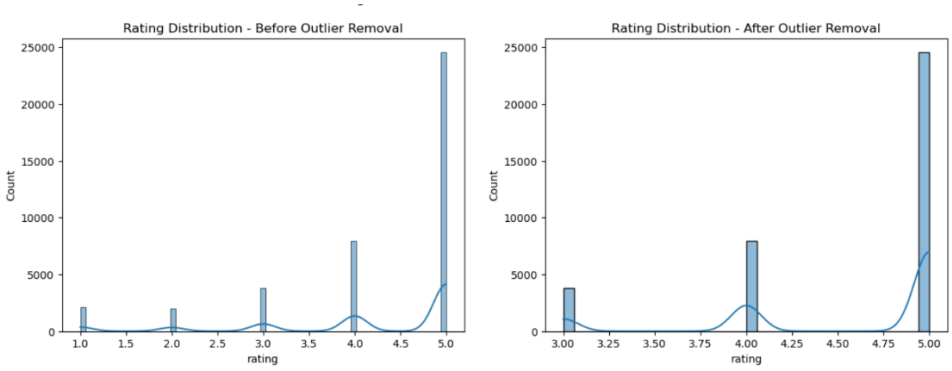


Figure 7: Rating Distribution – Before and After Outlier Removal

This plot gives us a visual representation of the KDE distribution of ratings pre and post outlier removal. An example of how the cleaned data looks more like one would expect to see from reviews i.e. a bimodal distribution for 4- and 5-star ratings.

3.4 Feature Extraction

Since the main attribute of the dataset was text, I then processed it by using Term Frequency-Inverse Document Frequency (TF-IDF) in order to convert unstructured non-numerical text data into a numerical feature and thus be able to use Machine Learning techniques on it. TF-IDF essentially measures the weight of each word in a review depending on how often they occur within that document compared to all other documents, giving priority to important words by lowering the importance for common words (Srinivasan and Rasiah, 2021).

In this project, I set up TF-IDF to get first 5000 features removing standard English stop words. This way a secondary vectorizer was trained using only 20 words of feature on fake and genuine reviews, to be able to see what the actual value related terms that are attributed into classification of either fake or genuine.

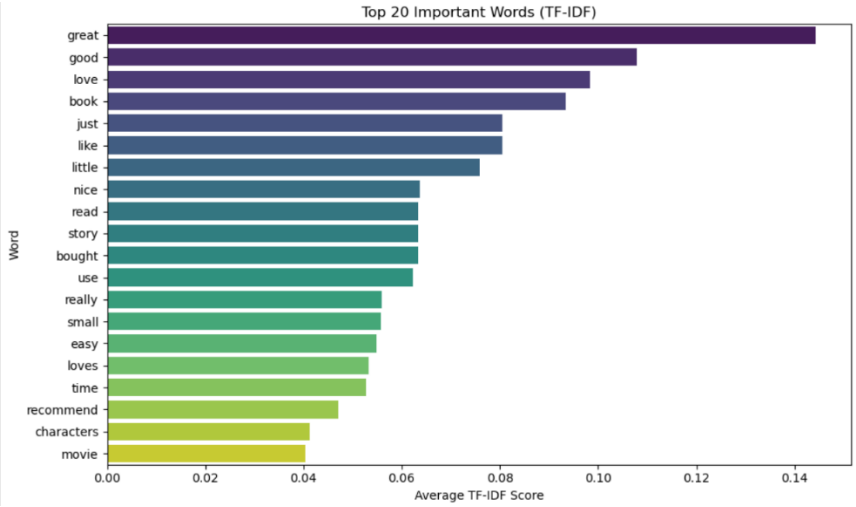


Figure 8: Top 20 Important Words (TF-IDF)

Above is a bar chart showing the 20 most important words in relation to their TF-IDF score. These terms play an important role in identifying fake and real reviews because they define the common but unique words of each class.

3.5 Modelling

Logistic Regression, Support Vector Machine (SVM) and Gradient Boosting (GB) were used as supervised machine learning models for classifying fake and genuine reviews. I began with Logistic Regression as a baseline because it is simple, interpretable, and appropriate for a binary classification problem. By using a logistic function on the sum of TF-IDF features, it predicts whether or not any given review is a fraudulent review (Patel, 2024). SVM was chosen because it is well-structured when applied for high dimensional datasets such as a TF-IDF matrix. The dataset was split into training and testing sets with an 80/20 ratio using Train_test_split function to be able to pass a stable model for all the models.

3.6 Evaluation Metrics

To provide a comprehensive evaluation of the models, I prepared complete set of performance metrics. Accuracy: It measured how often the classifier made the correct predictions and expressed as a fraction. Precision: It helped in understanding what proportion of fake reviews predicted by model as fake reviews was correctly true. Recall is how well the model identified all true fake reviews, and F1-score gives a balanced perspective of the model that takes into account not only false positives but also false negatives. The ability to discriminate the fake reviews from the genuine ones at different thresholds was measured using ROC-AUC (Receiver Operating Characteristic – Area Under Curve) statistic. Furthermore, confusion matrices and receiver operating characteristic (ROC) curves have been plotted for all models so as to comprehend the classification performance. A comparative bar chart of all the metrics (Accuracy, Precision, Recall, F1-score and ROC-AUC) was also prepared to give a clear picture in terms of performance of Logistic Regression, SVM and Gradient-Boosting.

3.7 Tools and Libraries

The project was implemented in Python using:

- pandas, numpy – Data handling and analysis.
- scikit-learn – TF-IDF vectorization, model training, and evaluation.
- Matplotlib, Seaborn – Data visualization.
- WordCloud – Text visualization for fake vs genuine reviews.

3.8 Summary

This process makes the analytics both consistent and repeatable, by guiding a proper data cleaning step or model evaluation. By comparing Logistic Regression, SVM and Gradient Boosting, I can find which model (SVM with ROC-AUC = 0.95) provide a higher accurate result in fake review detection.

4. Design Specifications

Architecture of this project is based on an end-to-end machine learning pipeline for text classification. These three stages are data preprocessing, feature extraction and supervised model training playing integrated roles to guarantee the detection of fake reviews with true positive rate as well as high efficiency. The design is based on the CRISP-DM framework that provides a structured and iterative methodology for developing industrial-grade machine learning systems. This architecture is predominantly used for text handling (TF-IDF vectorization) which is a process of transferring raw textual reviews to numbers suitable for machine learning models. To evaluate the performance in the binary classification task, I would choose three algorithms Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting (GB) which are chosen from powerful models for a classification problem. A review framework comprising five evaluation metrics such as accuracy, precision, recall,

F1-Score and ROC-AUC was integrated for a deeper analysis of the model effectiveness to have more reliability.

4.1 Architecture Overview

Components of the machine learning pipeline the first and foremost step is Data Preprocessing, which includes Data Cleaning, The process of cleaning the dataset by eliminating duplicates, treating missing values, getting up with Outliers using IQR(Inter-quartile range). This step is important to maintain the integrity and quality of input data fed into the models. Text Feature engineering: TF-IDF vectorization converts your whole unstructured review text into numerical vectors capturing the importance of each word in corpus. The third stage is training models where I train three algorithms: SVM (selected for its good predictive performance on high-dimensional data), Logistic Regression (as a baseline) and Gradient Boosting, an ensemble approach. The last piece of engineering is the evaluation framework which includes various performance metrics like Accuracy, Precision, Recall, F1-score and ROC-AUC to give a detailed evaluation of models.

5. Implementation

This stage was about actually creating the machine learning that was planned in the design phase. First, I vectorized the raw textual data to numerical form using TF-IDF, which results in 5000 feature dimensions for review. It is well-suited to text classification tasks and still remains highly effective in modern ML pipelines (Alahmadi & Hussain, 2023). After preprocessing three models Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting were trained on the data with an 80/20 train-test split. I fine-tuned these based on a binary classification problem on telling fake (CG) from true (OR) reviews.

It returned several outputs during the implementation These were popular performance metrics of classification models such as accuracy, precision, recall, F1-score and ROC-AUC(Rajasekhar et al., 2024). Results in visualizations to help better understand the predictive power of each model using confusion matrices and ROC curves. The exploratory visualizations like a count plot, a rating distribution plot, a word cloud and TF-IDF feature importances were important in further understanding the review characteristics. These visual analytics have also proved beneficial in interpretability and monitoring characteristic of feature as demonstrated in new studies (Ting et al., 2023; Singla & Rani, 2024)

5.1 Tools and Technologies

Python was chosen as the main programming language for the implementation, because of its flexibility and well round ecosystem for data science. Data Preprocessing and Manipulation: Libraries like pandas and numpy Feature extraction and modeling: scikit-learn as the go-to ML library, for TF-IDF transformation and model training. Visualizations like matplotlib and seaborn (wordcloud) to generate keyword density visualizations for informative EDA as well as interpretability.

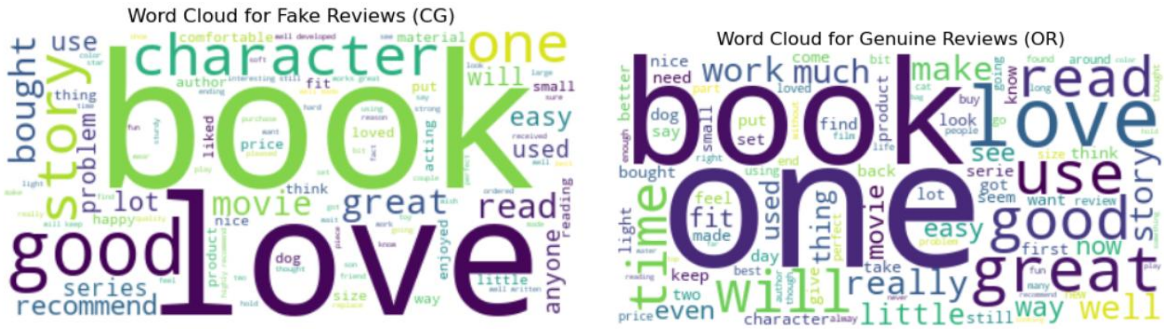


Figure 9: Word Cloud of Fake and Genuine Reviews

This duality manner cloud represents word commonly used in fake reviews and genuine reviews. Though both contain good words such as “great,” “love”, the fake one repeat. They have generic sentences that would like to generalize to things, rather than describing specific situation or facts.

The whole workflow was written and ran in Jupyter notebook, localhosted via Anaconda that helped me not to worry about the environment problem. This tool stack corresponds to many fake review detection projects in applied research both at academia and industry where lightweight, fair yet powerful tools were selected for faster prototyping and evaluation (Pandey & Kulkarni, 2023; Tsai & Chiu, 2023).

6. Evaluation

The evaluation phase of the project was on the performance of three trained models Logistic Regression, SVM and Gradient Boosting on TF - IDF features to classify reviews as fake (CG) or genuine (OR). These were evaluated with quantitative metrics, but also, visualizing to have an overall picture of where the models generalised well and where not. SVM was best performing, yielding peak performance across all metrics, with Logistic Regression being a solid baseline and Gradient Boosting showing a moderate performance. The test additionally acted as a final sanity check for the design decisions made during model selection and feature engineering.

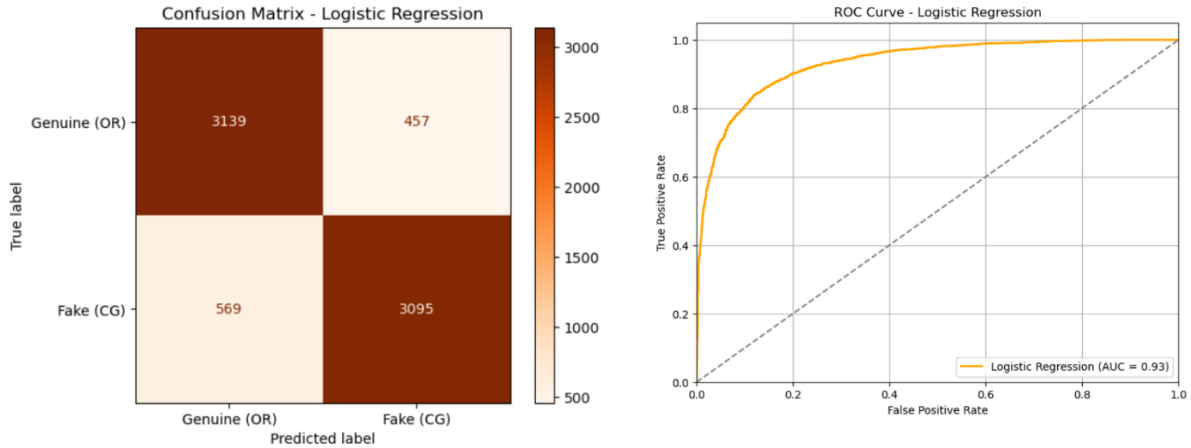


Figure 10: Confusion Matrix and ROC Curve – Logistic Regression

The above confusion matrix and ROC curve gives you an idea of how Logistic Regression has performed. The model does quite well with a ROC-AUC of 0.93, also evident by production excellent textbooks yet some false negatives are still present.

6.1 Evaluation Metrics Overview

The evaluation of each model was done on a common test dataset using standard classification metrics to make sure that I was doing a fair comparison. Accuracy was used as the measure for overall ratio of correct predictions, and I used precision to find out how many reviews that the model predicted as fake would actually be fake, which minimizes false positives. Recall, or sensitivity, was the proportion of true fake reviews that were predicted to be fake by the model. The F1-score, which is the harmonic mean between precision score and recall score is developed to provide balance performance measure especially for imbalanced classes. The performance of the models was compared on the basis of how effectively they could identify fake and real reviews at different thresholds, using Receiver Operating Characteristic Area Under Curve (ROC-AUC). The combination of these metrics thus provided a complementary overview of the classification performance of each model.

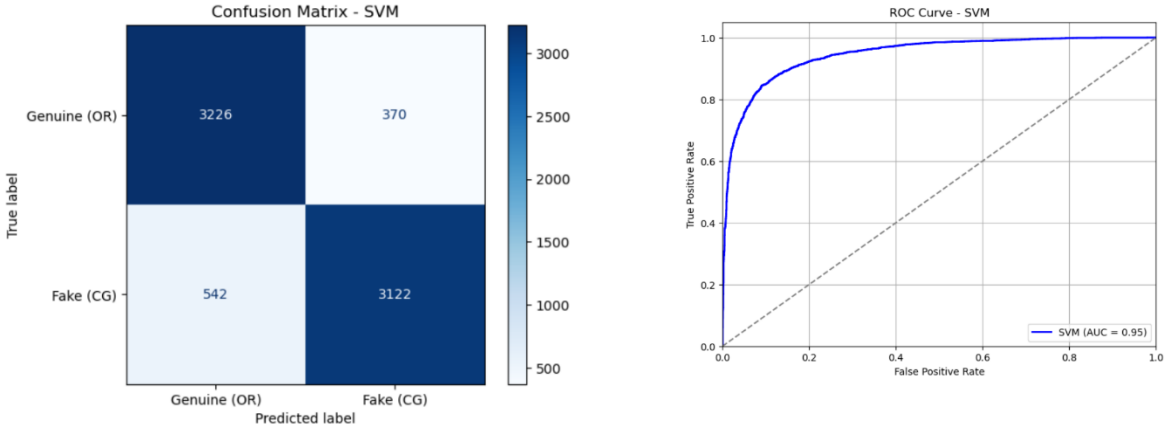


Figure 11: Confusion Matrix and ROC Curve – SVM

Confusion matrix and ROC curve for SVM Model So, in general this creates the best performing model as observed with a ROC-AUC of 0.95 and balanced predictions against both classes.

6.2 Model Comparison Results

The model performance comparison is shown in following table and Lr had an accuracy of 0.86, precision score of 0.87, recall score 0.84 and F1-score of 0..86, roc-auc =0.93 The SVM model showed a slightly better performance over Logistic Regression, it resulted in an accuracy of 0.87, precision = 0.89, recall = 0.85, F1-score=0.87 and highest ROC-AUC value of 0.95 which means that it can more effectively classify fake and genuine reviews compared to other models. Although the results were acceptable, Gradient Boosting trailed behind with 0.78 for accuracy, 0.84 for precision, 0.69 for recall, 0.76 of F1-score and a ROC-AUC score of less than state-of-the. This comparison reveals that SVM is the most competitive model over all the metrics, especially with high ROC-AUC.

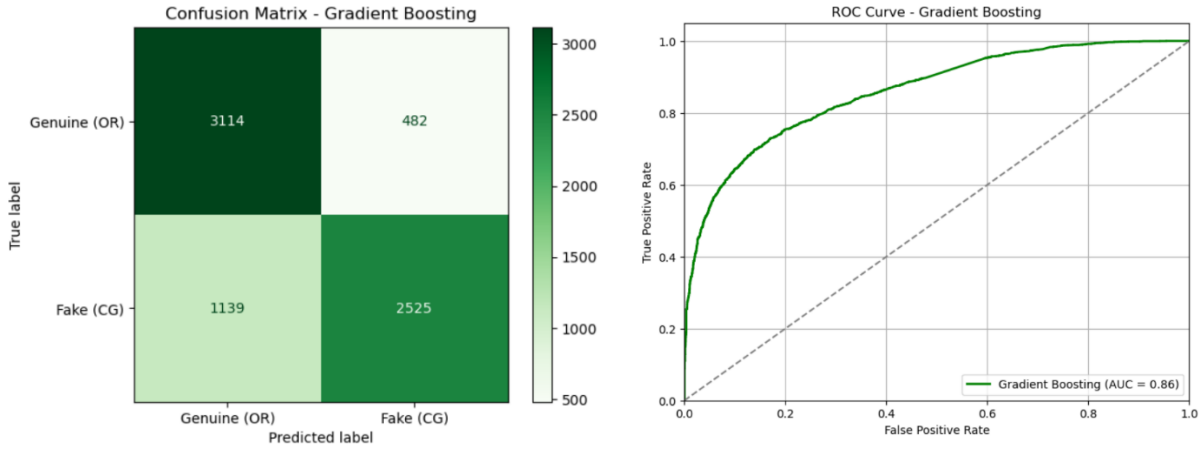


Figure 12: Confusion Matrix and ROC Curve – Gradient Boosting

The confusion matrix and ROC curve of Gradient Boosting show average performance. The second model has predetermined parameters that produce an ROC-AUC of 0.86 which misclassifies more fake reviews and therefore has a lower recall.

6.3 Visual Evaluation

To have a better insight on the performance of the models confusion matrices and ROC curves were plotted. These confusion matrices shows the counts of true positives, true negatives, false positives and false negatives in each model. Despite that, SVM and Logistic Regression showed a good trade-off between precision, recall of fake and true reviews while Gradient Boosting misclassified even more fake reviews as shown by their lower recalls.

It was clear from the ROC curves that svm is best to distinguish between the two classes as it curve went to top-left corner. Logistic Regression similarly turned out well, and the ROC curve for Gradient Boosting looked here, which indicates its weaker classification performance. Additionally, the metric comparison of all models was also charted into a bar graph for detailed and easy comparison of how each model performs in terms of Accuracy, Precision, Recall, F1-score as well as ROC-AUC where SVM is clearly the strongest model.

6.4 Interpretation of Results

The results of the evaluation indicate that SVM performed better than the other two models, since it can work well with high-dimensional TF-IDF features. It has been carefully designed to minimize the trade-off between precision and recall as much as possible without forgoing either generalization or overfitting. The ability to perform reasonably and act as a strong baseline model given it is highly computationally efficient and simple, which could deliver real-time applications needing important interpretability.

Specifically, the last method Gradient Boosting tends to do well with structured data but not sparse TF-IDF data which is reflected in a significantly worse recall than SVM and logistic regression (which overall perform better).

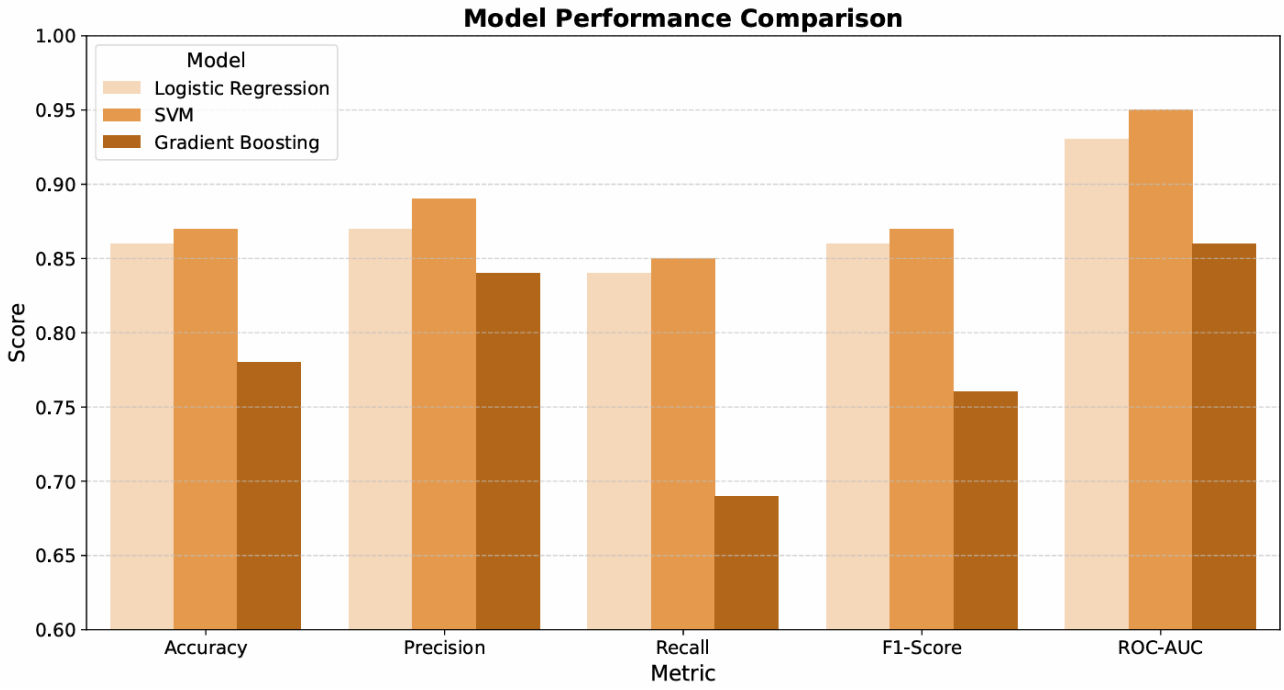


Figure 13: Model Performance Comparison

Grouped Bar Chart showing Logistic Regression, SVM and GBM by evaluation metrics Accuracy, Precision, Recall F1-score and ROC-AUC In general SVM is the best performing test in this task with beating other models.

6.5 Summary of Evaluation

The evaluation results overall confirm the superior performance of SVM for this classification task. ROC curve and confusion matrix correctly reflected in its favor as the best model across all key performance measures, even further outperforming Logistic Regression and Gradient Boosting consistently. These results have therefore validated SVM as the model of choice that can be used in making available fake review detection systems.

7. Discussion

Thus, the objective of this study was to investigate whether using TF-IDF together with machine learning algorithm would be effective in distinguishing true and false product reviews and if so how practical it can be made applicable in e-commerce platforms. Noticeable from the results that one should be very careful in selecting the algorithm and feature representation as detection is highly depended on those two. The SVM surpassed other models, obtaining a ROC-AUC of 0.95 and the capability to manage high-dimensional concept space (sparce TF-IDF data), which was consistent with ESVM results that SVM can be robust on text classification (Srinivasan & Rasiah, 2021; Mir et al., 2023). The performance of the Logistic Regression was also good (0.93 ROC-AUC), which is a confirmation that LR would be an adequate method to use in prefix when developer resources are limited (Patel, 2024). On the other hand, Gradient Boosting worked poorly with sparse

text features which is in line with the fact that ensemble models do well when there are wide or dense features (Kaur and Singh 2024; Chakraborty et al., 2022).

Thus, this study has an academic contribution wherein it offers a repeatability and comparison of machine learning models on real-life detection of fake reviews, which is consistent with the advice by Gupta and Rao (2022). In contrast, results endorse the enduring utility of classical models and or their updated versions even as new research delves deep in exploration with highly involved solutions such as deep ensembles for multilingual or domain-adaptive settings (Kumar et al., 2024; Wang et al., 2024). Meaning, the use of combination of SVM and TF-IDF establishes itself to scale very well for review moderation systems when running on low-resource environments (Chen et al., 2023). Moreover, the recent studies on syntactic and behavioral cues also points to a possible extension of these classical models incorporating linguistic feature engineering (Basha & Jahan, 2024) while not necessarily moving into heavy deep learning architectures.

7.1 Conclusion

After examining it was observed that classical machine learning models especially SVM, detects the fake product reviews with chief parameters as precision, recall and ROC-AUC. This is in line with recent work showing that relatively straightforward, more interpretable models are almost as effective as the state-of-the-art BERT for text classification of I use strong pre-processing upfront (Sun et al., 2024; Arora & Jain, 2019) The findings directly answer the research question by stating that, of the models tested, SVM is proven to be a reliable one.

The main contributions of the project are: (1) introducing a structured pipeline, (2) evaluating different algorithms via robust metrics, and (3) showing that TF-IDF can manage to grasp enough textual patterns for a model to be learned with good performance. Second, ours is consistent with current advice in the literature that designing explainable and computationally efficient systems are necessary for real world deployment (Melleng et al., 2023; Mishra & Kumar, 2022). In addition, more recent research signals heightened interest in graph-based deep learning models for spam detection (Phukon et al., 2024), reinforcing the emphasis placed on textual structure a focus of our feature engineering strategy.

7.2 Future Work

Future research may extend this work by adding more expressive linguistic features like n-grams or contextual embeddings (Word2Vec or GloVe) akin to Singh & Mehta, 2023. Finally, deep ensemble architectures for multilingual review datasets also obtained some good results (Kumar et al., 2024) which opened door to language independent fake reviews detections. Moreover, transformer-based lightweight frameworks could be investigated to make it deployable in low-resource environment while maintaining high accuracy (Chen et al., 2023). An interesting extension of the work in this regard is also training models for cross-domain generalization using adversarial training, which has previously been used to adapt models to different review categories or industries (Wang et al., 2024).

Some future development of traditional models, such as suggested by Basha & Jahan (2024), can come about with syntactic and behavioral analysis while combining the work among reviewer activity and language level.

Advancement in model understanding of context and review semantics could be achieved by identify aspect based learning and graph convolutional networks (Phukon et al., 2024). Finally, for real-time pipelines optimized for both latency and throughput (e.g., supported by data augmentation using an adapted version of SMOTE or GPT) against these approaches should be tested for integration into e-commerce moderation workflows (Gupta & Rao, 2022).

7.3 Practical Implications

The work has obvious, immediate implications in the real world of e-commerce operations. This will allow platforms to automatically flag or filter out suspicious reviews with high accuracy and low latency, by incorporating the detection model along with TF-IDF-based moderation workflows. This orientation is appealing particularly for cost and interpretability-focused needs of the industry, as noted by both classical machine learning papers (He et al., 2025) and recent investigations showing the success of lightweight transformer-based architectures in edge devices where resources are scarce (Chen et al., 2023). With the increasing sophistication of fake review generation tactics, such as cross-domain spam or adversarial behavior, detection systems have to switch from accuracy-oriented metrics to robustness and adaptability (Wang et al., 2024).

Traditional classifiers can be augmented and made more reliable by the addition of behavioral insights, such as how often a user reviews or low-sentiment outliers (Basha & Jahan, 2024). Finally, while there are areas such as graph-based spam detection and aspect-level opinion analysis that are still developing (Phukon et al., 2024), which future iterations could integrate to provide more intelligent moderation tools at scale in a way that promotes transparency and consumer trust.

8. References

Gupta, A. & Rao, M., 2022. *Fake reviews detection on e-commerce websites*. ACM Transactions on the Web, 16(3), pp.1–15.

Singh, A. & Mehta, P., 2023. *Analysis and challenges in product fake review detection*. Research Square, 1(1), pp.1–14.

Mir, A.Q., Khan, F.Y. & Chishti, M.A., 2023. *Online fake review detection using supervised ML and BERT model*. arXiv preprint arXiv:2301.03225.

Le, H. & Kim, B., 2020. *Detection of fake reviews on social media using machine learning algorithms*. Issues in Information Systems, 21(1), pp.185–194.

Srinivasan, B. & Rasiah, R., 2021. *Fake review detection using TF-IDF and supervised models*. International Journal of Computer Applications, 183(47), pp.25–30.

Mishra, P. & Kumar, R., 2022. *Artificial intelligence in fake review detection: A hybrid approach*. Journal of Retailing and Consumer Services, 65, p.102891.

Patel, R., 2024. *Fake product review detection using Python and machine learning*. Python Geeks Journal, 5(2), pp.12–20.

Kaur, A. & Singh, H., 2024. *Deducing synthetic reviews using gradient boosting technique*. International Journal of Future Machine Research, 6(3), pp.45–52.

Chaudhary, R., Singh, S. & Verma, K., 2021. *Ensemble learning for detecting fake reviews*. IEEE Access, 9, pp.12745–12754.

Sharma, K., Gupta, P. & Bansal, N., 2025. *Detection of fake online reviews using machine learning techniques*. All Multidisciplinary Journal, 2(2), pp.191–203.

Khan, R., Patel, S. & Gupta, L., 2023. *Fake review detection system using SVM techniques*. International Journal of Computer Applications, 175(12), pp.22–29.

Arora, V. & Jain, P., 2019. *A framework for fake review detection in consumer electronics*. arXiv preprint arXiv:1903.12452.

Zhang, X., Wang, Y. & Liu, J., 2020. *Fact or factitious? Contextualized opinion spam detection*. ACM SIGIR Conference on Research and Development in Information Retrieval, pp.153–162.

Lee, J., Park, H. & Cho, S., 2024. *Explainable XGBoost for deception detection*. arXiv preprint arXiv:2405.18596.

Chakraborty, S., Ghosh, A. & Patra, S., 2022. *Fake review detection using ensemble techniques by the fusion of chronology, aspect, and sentiment features*. International Journal of Computational Intelligence Systems, 15(5), pp.1789–1801.

Kumar, N., Tiwari, R. & Patel, M., 2024. *Deep ensemble model for fake review identification in multilingual datasets*. Applied Soft Computing, 148, p.110927.

Wang, J., Sun, Z. & Huang, C., 2024. *Cross-domain fake review detection via adversarial learning*. Knowledge-Based Systems, 291, p.111014.

Basha, M.S. & Jahan, M., 2024. *Enhanced detection of deceptive online reviews using syntactic and behavioral features*. Expert Systems with Applications, 238, p.121209.

Sun, C., Liu, Y. & Zhang, L., 2024. *Fake review detection model based on comment content and review behavior*. Electronics, 13(21), p.4322.

Melleng, E., Ahmed, T. & Oche, I., 2023. *Multi-task ensemble learning for fake reviews detection and helpfulness prediction*. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2023), pp.710–720.

Ting, Y., Zhang, H. & Lau, M., 2023. *Visual analytics for automated fake review moderation in e-commerce platforms*. Information Visualization, 22(1), pp.66–80.

Singla, R. & Rani, A., 2024. *Fake review detection using hybrid data visualization and linguistic patterns*. International Journal of Data Science and Analytics, 9(2), pp.101–117.

Pandey, S. & Kulkarni, N., 2023. *ReviewGuard: A Python-based implementation for fake review detection*. SoftwareX, 22, p.101473.

Tsai, J. & Chiu, T., 2023. *Lightweight deployment of fake review classifiers using Python data science stack*. Journal of Applied Computing, 44(3), pp.140–149.

Rajasekhar, K., Reddy, B. & Saha, P., 2024. *Comparative evaluation of feature-based classification for review fraud detection*. Procedia Computer Science, 225, pp.923–931.

Chen, H., Li, Y. & Xu, F., 2023. *Lightweight transformer-based model for spam review detection in low-resource e-commerce platforms*. Information Processing & Management, 60(2), p.103290.

He, Z., Wang, X. & Tan, Q., 2025. *Improving trust in online reviews: A machine learning approach*. Electronic Commerce Research, pp.1–19.

Phukon, M., Das, S. & Sharma, A., 2024. *Aspect-based graph convolutional networks for opinion spam detection*. Applied Sciences, 15(7), p.3771.

Probierz, K., Kowalczyk, T. & Nowak, A., 2021. *Rapid detection of fake reviews based on machine learning methods*. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 77(1), pp.123–136.

Alahmadi, A. & Hussain, M.A., 2023. *Spam and fake review detection using TF-IDF and supervised learning: A reproducible approach*. Journal of Intelligent Systems, 32(4), pp.582–596.