ML 4641 Group 49 Final Report

Introduction

Online reviews strongly influence consumer decisions, with over 80% of Americans consulting them before purchases [5]. However, fake reviews can mislead buyers, damage business reputations, and cause financial loss. To address this, we plan to train a machine learning model to detect fabricated reviews. We will use a dataset introduced in prior research [5], which compared Transformer models with an SVM hybrid approach. The dataset (Fake Reviews Dataset) contains **40,412** product reviews labeled as either computer-generated or real, plus features such as review ratings and product categories. The data is roughly balanced between genuine and fake reviews, making it well-suited for supervised learning.

Problem Definition

E-commerce platforms face growing difficulty in distinguishing authentic reviews from fraudulent ones. Fake reviews can distort product ratings, mislead consumers, and erode trust in online marketplaces, while also giving dishonest sellers an unfair advantage. Manual moderation is sometimes used but is inefficient and prone to error. To address this, we propose developing an automated system to identify fake reviews. Such a model would help consumers make informed decisions based on reliable ratings and allow businesses to forecast sales more accurately. By improving transparency and reducing manipulation, this system could promote fair competition and strengthen long-term customer confidence in e-commerce platforms.

Methods

Data Preprocessing

For data processing, we implemented a structured pipeline to make sure that the dataset was clean and balanced for training. The pipeline first gets the raw dataset and cleans the data by normalizing all the text to be lowercase, removing any punctuation or white space, and dropping the rows that have invalid labels. We also made sure to get rid of any repeated reviews based on whether the content had identical text so that the data leakage between splits wouldn't happen.

After we finished cleaning, we analyzed the distribution of labels to ensure class balance, and then our team created derived features like length of review to help later in model interpretability. Lastly,

we created stratified train, validation, and test splits with a ratio of 70/15/15 so that we can preserve the label proportions across all the subsets. These splits were then saved as separate csv files so that we can consistently reuse them across various experiments.

We chose this structured approach because it allows for data integrity, reproducibility, and fairness for model evaluation. By modularizing the preprocessing and automating the file generations, our team was able to reduce human error and make sure that all experiments were consistent, aligning all workflow with machine learning best practices.

Since Term Frequency-Inverse Document Frequency (TF-IDF) has shown to be effective for fake review detection in Nhut-Lam Nguyen (Nguyen, 2023) [1], TF-IDF was implemented to convert the textual reviews into numerical vectors that the machine learning models could use. It gives higher weight to words that are unique to a review and less to a more common word like “the” and “and”, helping the model focus on more meaningful words. Using the vectorizer TfidfVectorizer with unigrams and bigrams, the training data was transformed into valid sparse matrices. Then these matrices were saved in the reports/ folder for the next step.

Model Selection and Configuration

We implemented the following supervised learning algorithms to classify each review: logistic regression, decision trees, random forest, and support vector machines.

Logistic Regression

We chose to implement logistic regression (supervised method), because we wanted to use soft assignments for greater flexibility. This is an effective solution because by examining the coefficient weights, we can identify which words are the strongest predictors. Furthermore, its effectiveness is well established in literature on fake review detection [3]. This model is also very fast to train, O(ndi) where n = number of reviews, d = number of features, and i = number of iterations. In addition, Scikit-learn has its own LogisticRegression class.

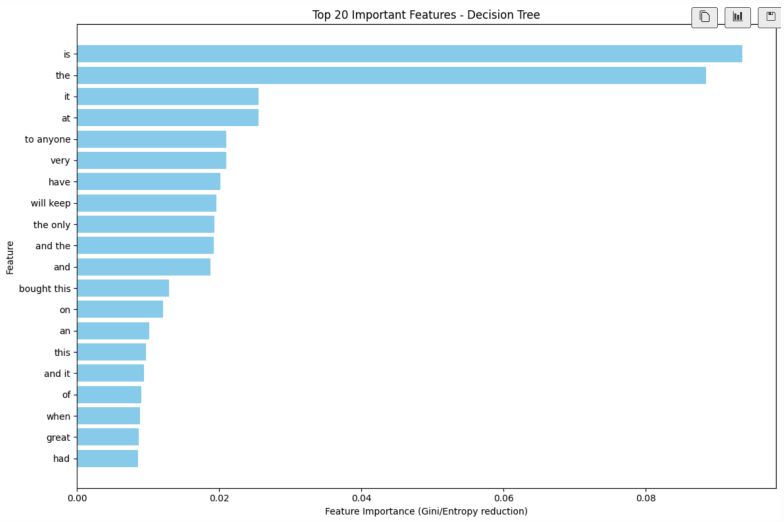
For model optimization, we experimented with multiple regularization strengths (C values) and selected **C = 10** based on validation performance. We applied stratified 70/15/15 train–validation–test splits to preserve class balance and reproducibility. To ensure fair evaluation, we standardized preprocessing across all models and used a consistent random seed. Evaluation was performed on held-out data using `sklearn.metrics` to compute Accuracy, Precision, Recall, F1, and Balanced Accuracy.

To optimize the trade-off between false positives and false negatives, we tuned the decision threshold instead of relying solely on the default 0.5. We adjusted the probability cutoff until achieving a precision target of ≥ 0.90, resulting in a threshold of **0.279**. This modification improved model precision while maintaining strong recall, demonstrating the flexibility of logistic regression for calibration.

Decision Tree

We implemented decision trees, which classify through a sequence of feature-based questions, following learned rules to make hard class assignments. This is an effective model because fake review detection is a task where model interpretability is essential to supporting the final decision, and it is the most interpretable because we can clearly follow its decision path [3]. Also, our dataset includes other features like rating, which decision trees can easily consider both the rating and text reviews together. Furthermore, scikit-learn has its own DecisionTreeClassifier class, making it easy for us to create and train the model.

After training our initial Decision Tree model, we observed that it struggled with the high-dimensional TF-IDF feature space (125,334 features), showing signs of overfitting as will be discussed in the Results and Discussion section. To address this, we implemented Chi-squared feature selection as a dimensionality reduction technique. Chi-squared feature selection is a supervised method that tests the statistical independence between each feature and the class label, selecting the top k features that show the strongest association with fake/real classification. We chose this method because it is computationally efficient for high-dimensional sparse data, preserves class-discriminative information, and is interpretable—allowing us to identify which words and n-grams are most important for distinguishing fake from real reviews. We selected the top 10,000 features based on chi-squared scores, reducing dimensionality by 92% while maintaining the most relevant features for classification.

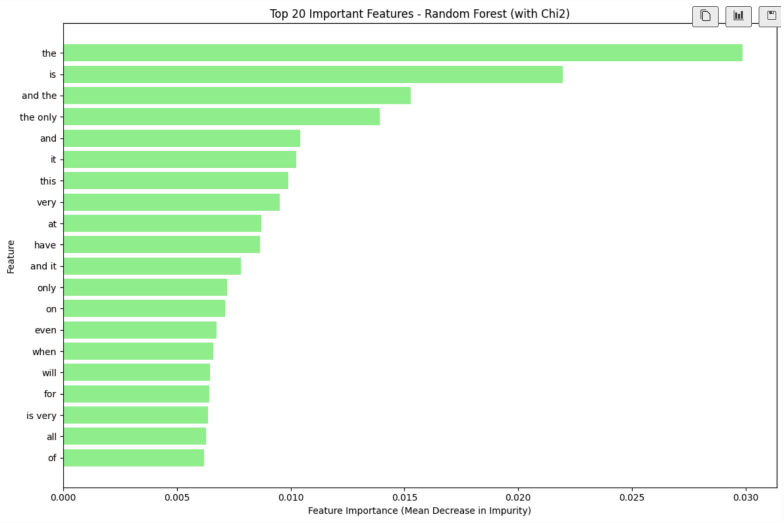


Decision Tree Visualization

Random Forest

Following the Decision Tree with Chi-squared feature selection, we implemented Random Forest as an ensemble method. Random Forest addresses the limitations of a single Decision Tree by constructing multiple diverse trees, where each tree sees different data samples and different feature subsets. This ensemble approach reduces overfitting, improves generalization, and can

better capture complex patterns in the data. We chose to implement Random Forest based on the results from Decision Tree with Chi-squared, which will be discussed in the Results and Discussion section. Random Forest is particularly effective for high-dimensional text data because feature subset sampling works well with the chi-squared selected features, and the ensemble voting mechanism helps overcome the systematic errors we observed in the single Decision Tree model.



Random Forest Visualization

Support Vector Machine (RBF Kernel)

To model fake review detection in a high-dimensional TF-IDF feature space, we implemented a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. SVMs are well-suited for text classification because they create non-linear decision boundaries and perform effectively on sparse, high-dimensional data. The RBF kernel was chosen because it captures complex relationships between word features that linear models (e.g., Logistic Regression) may miss.

We performed grid search with **3-fold cross-validation** over the following parameter grid:

- **C ∈ {0.1, 1, 10}**
- **γ ∈ {"scale", 0.1, 0.01}**

This resulted in **27 total fits**. The best-performing model was:

`C = 10 gamma = 0.1 Best CV F1 = 0.9409 Total training + tuning time ≈ 1645 seconds`

This higher-capacity SVM (larger C and γ) suggests that the dataset benefits from a more flexible, expressive boundary.

Instead of using the default decision threshold, we optimized the cutoff using the **precision-recall curve on the validation set**, selecting the threshold that maximized F1.

Optimal threshold on validation set:

- **Threshold:** 0.0198
- **Precision:** 0.9544
- **Recall:** 0.9468
- **F1:** 0.9506

Results and Discussion

To assess our models, we used various quantitative metrics mentioned in [2]. Accuracy indicates the overall percentage of correctly classified reviews, whereas balanced accuracy addresses potential class imbalance by averaging recall between real and fake reviews. To emphasize error sensitivity further, we present the F1 score, which reconciles precision (the proportion of reviews identified as fake that are indeed fake) and recall (the proportion of all fake reviews accurately recognized). The objective of our project is to develop a detection system that is not only precise but also computationally sustainable and ethically sound. Specifically, we strive to reduce false positives, thereby safeguarding genuine users from being unjustly flagged, while still ensuring robust detection of fraudulent activities.

Logistic Regression Results

1. Performance Interpretation

Our baseline Logistic Regression model achieved strong generalization without overfitting, as shown in the table below. The slight drop between validation and test suggests minimal variance, implying stable model behavior on unseen data.

Metric	Validation	Test
Accuracy	94.77%	94.26%
F1-Score	0.9473	0.9420

2. Precision–Recall Trade-off

Increasing the decision threshold from the default 0.5 to 0.279 enhanced precision while preserving high recall. This is particularly valuable in fake review detection, where false positives can erode user trust. Our results show that precision can be improved with minimal F1 loss, aligning with our ethical goal of minimizing harm to legitimate users.

Threshold	Precision	Recall
0.5 (Default)	0.9538	0.9409
0.279 (Optimized)	0.9001	0.9762

3. Interpretability

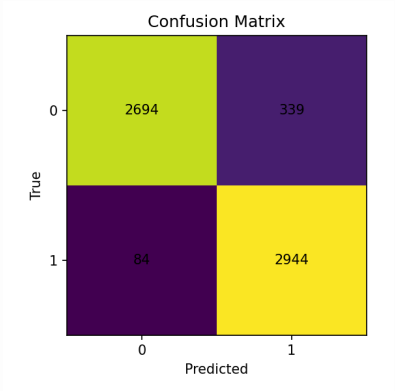
Logistic Regression's linear weights allow direct interpretation of the most influential features. For example, tokens such as *"amazing," "highly recommend,"* or *"worst"* may exhibit strong positive or negative coefficients. This interpretability supports explainability in moderation systems, a critical factor for transparent and fair AI.

4. Computational Efficiency

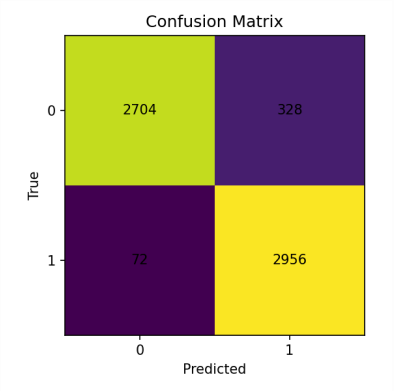
Model training completed within seconds on CPU due to the efficiency of linear solvers on sparse TF-IDF matrices. This suggests that our baseline is scalable to larger datasets and suitable for deployment in real-time moderation environments.

5. Limitations and Next Steps

In order to further explore hindrances in our model, we computed some more metrics. Firstly, we visualized a confusion matrix (on our test and validation data) of our model's predicted and true output values:



Test Data



Validation Data

In both confusion matrices, there appears to be more false negatives than false positives (top right cell). This hints at nuances in the dataset which helps capture the blurry distinctions between fake and real reviews. It seems that the model is generally more confident in a review being fake, given the larger number of false negatives. This hints to a limitation of our classification model at capturing such nuances. This is also natural since our threshold value is lower than 0.5, at 0.279.

This would naturally lead the model to be more eager to classify a review as fake. We did set the threshold as such to avoid false positives, however, which we did relatively achieve.
We also performed an F1 and balanced accuracy analysis on various slices on the test data. We essentially indexed specific values of each of our features and performed a metric analysis on those outputs with the model's predicted outputs. We did this for each of the ratings, categories and the length of the review. We grouped the lengths into three categories: long, medium and short, by using Panda's qcut method. The results of this analysis are as follows (ba is balanced accuracy):

Feature	Feature Slice Value	F1	Balanced Accuracy	Count
rating	5.0	0.9342	0.9345	3724
rating	4.0	0.9466	0.9477	1151
rating	2.0	0.9505	0.952	314
rating	3.0	0.9671	0.9666	570
rating	1.0	0.9655	0.9665	302
category	Toys_and_Games_5	0.9396	0.9398	593
category	Sports_and_Outdoors_5	0.9437	0.9452	582
category	Tools_and_Home_Improvement_5	0.9591	0.9579	499
category	Kindle_Store_5	0.9317	0.9339	754
category	Clothing_Shoes_and_Jewelry_5	0.9362	0.9405	603
category	Electronics_5	0.9617	0.9617	596
category	Pet_Supplies_5	0.9495	0.9478	671
category	Books_5	0.9455	0.9459	628
category	Home_and_Kitchen_5	0.942	0.9431	616
category	Movies_and_TV_5	0.9091	0.9112	519
review length	short	0.8931	0.8901	2104

Feature	Feature Slice Value	F1	Balanced Accuracy	Count
review length	medium	0.9549	0.9549	1965
review length	long	0.9869	0.9872	1992

Note that the sum of the count for each group of features adds up to the total data points in the test data (around 6061). Immediately off the bat, and perhaps the most obvious, are the lower metric values for the short review length. This is expected since there is only so much information that can be encoded in fewer words. The model could do a better job capturing patterns in these smaller sized reviews, since false negatives are more likely here. We want to avoid ignoring useful reviews, despite their length. We see that the metric is high for long reviews, which somewhat confirms the notion that lengthier reviews provide the model with more information leading to more accurate classifications. We can see that in terms of ratings, the reviews with 5.0 stars had slightly lower metric scores than the other ratings. This may be because high-rating reviews may look more generic, making it harder for the model to distinguish the real and fake reviews. Interestingly, it doesn't have that issue with 1.0 star reviews, which suggests that the reviews aren't as generic. This may hint at possible updates to our dataset, in order to ensure our model can identify nuances to help distinguish between real but somewhat generic reviews and fake ones. Now, the variation of the metric scores between categories appears to be somewhat arbitrary. It does seem that movies and tv reviews have more subtle fake reviews or contain unique writing patterns that is difficult for the model to pick up on. Finally, we viewed the top 20 false negatives and false positives from our data and saved them in a table:

Top 20 False Negatives:

Reviews	True	Predicted	Predicted Probability
good package but i hardly noticed it until the night before i got home from work	1	0	0.02164
excellent product and a much better quality than the one you get at walmart for 50	1	0	0.0226
didn t quite sit flat on the table but it worked perfectly i just got this one	1	0	0.03225

Reviews	True	Predicted	Predicted Probability
really great iron my only thing is it doesn t stay in place after washing it slides out	1	0	0.03834
twisted premise check dark check dark check dark check dark check dark check dark check dark check dark check dark check dark check dark check dark check	1	0	0.04259
item of exceptionally excellent quality and no more than a	1	0	0.05976
u wont be disapointed blueray no more wonderful movie i recommend you boring it as a film no way can i not	1	0	0.06437
so useful when shooting in low light and with very bright lighting if you have a full sized tripod you may want to look at the video or	1	0	0.07189
like this nozzle big improvement over the original very bright works great for the price	1	0	0.07327
this is a fantastic educational tool for the child it is easy for a child to understand there are several different songs including the kids song be happy and be happy and there are many different songs that they can sing as well there are also different variations on the songs	1	0	0.07444
delivered fast fits well looks good and feels comfortable	1	0	0.0755
seki city what more is there to say the price is nice the product is awesome	1	0	0.07834
my wireless connection drops from about 20 feet to about 15 feet this is a major issue if you are trying to run a wireless network through a wall you should go to the router and look at the bluetooth device otherwise if you are using the device for the web	1	0	0.0878

Reviews	True	Predicted	Predicted Probability
browsing it s a huge issue if you have a router that is really capable of handling			
this book was hot	1	0	0.08833
great product boils water and takes the time to drink out of	1	0	0.09002
i use this as a training tool and as a training tool when i first started using this	1	0	0.0905
great knife i m not used to carrying knives with long handles but this one is just perfect	1	0	0.09082
exactly as i wanted can t say enough good things about this tool	1	0	0.10099
perfect to throw in my kid s birthday cake and make them all love it i would recommend	1	0	0.10518
good western romance anna and jack have been through hell together but now they need each other can anna and jack find their way back together and get through the trials and tribulations that were causing	1	0	0.10922

There are various examples here where we can see where the model may have made misinterpretations. There are several reviews where words are repeated several times for emphasis. Most real reviews may not have such a pattern, and so the model likely saw the review as fake. Furthermore, some of these reviews use internet slang, which are technically grammatically

incorrect. The model thus may have classified the reviews as fake. There is also a fact that sometimes people write reviews that are somewhat “robotic”, and so more nuance is necessary when interpreting them. These might simply expose the flaws of the classification model. It is interesting how the probabilties aren’t on the border of the threshold but fairly far from it. The model was very sure of its predictions.

Top 20 False Positives:

Reviews	True	Predicted	Predicted Probability
the quality is great the color is awesome and the fit is perfect	0	1	0.98851
is s good strong suction vacuume world s good an the carpet also	0	1	0.98385
bought this for my wife and she was surprised to see how small it was had to return it	0	1	0.97226
my son is happy with it a very comfortable fit and the quality is durable to	0	1	0.97062
i love the paranormal and kinetic behavior the mind is very strong if you have the ability this was a good read would like to read more	0	1	0.96022
fits my 140 pound german shepherd not an easy thing to do	0	1	0.95299
i ve been using these for a few weeks now and i have had no problems the build quality is very good on these	0	1	0.95102
i love this series the characters are well developed i fell in love with both the male leads this book is a must read you will not feel cheated	0	1	0.94495
the slippers i bought my granddaughter were too small the exchange could not have been any easier sophie loves the look and feel of the shoe	0	1	0.94197

Reviews	True	Predicted	Predicted Probability
this standalone story was very good there were a lot of typos and misspellings but it was still worth reading	0	1	0.93723
this shirt fits my husband perfect and the quality is great he is typically an xl in everything and the xl fits him great	0	1	0.93692
dvd is great movie to watch and will love it	0	1	0.93227
we have had this for years and the kids still love to hop on it	0	1	0.9283
my husband had this in vhs and i decided to get in blue ray this is the only movie i have ever seen him watch repeatedly	0	1	0.92563
this is the second book in the series and its just as delightful as the first if you love the young adult genre but also love fairytales this one will definitely be up your alley the characters are fun the story will keep you engaged and the drawings at the beginning of each chapter are just lovely you will not be disappointed	0	1	0.91792
i received this fruit bowl at a discounted price in exchange for an honest review with saying that i am very impressed with the quality of this bowl and it s a lot bigger than i thought it was going to be which is great the stainless steel matches my kitchen flawlessly and holds quiet a bit of fruit i would highly recommend	0	1	0.91177
ainsley finds out more about herself she is going to pick the next alpha she just doesn t know who it will be this story is well written and has great characters i am ready for the next book to find out what happens	0	1	0.91044

Reviews	True	Predicted	Predicted Probability
great read kept move at a very good pace it was very easy to visualize the canyon walls as they climbed i would recommend this to anyone who would like to go and visit texas s high desert keep writing books of the caliber and i will keep reading them	0	1	0.90744
i ordered this for my husband he loves the set especially the noodle strainer	0	1	0.90138
they fit perfect they look expensive they are the most confortable shoes that i had ever i love the design	0	1	0.89016

Many of these reviews have typos which is something that many real reviews have. As such, the model may have picked up on this and thought these reviews were real. The issue is that the line between human typos and fake typos is hard to draw for a human, let alone a classification model. The model seems to have a hard time picking up on these nuances or rather when its real and when its fake. Some of the reviews in the bottom sound somewhat “robotic”, but again they are also reviews that a person may possibly write. Yet again, the predicted values are high, indicating that the model was sure of these predictions. It may be possible that the model doesn’t recognize the mentioned patterns effectively or correctly too. All in all, there are clear limitations in what the model is able to pick up on through these reviews.

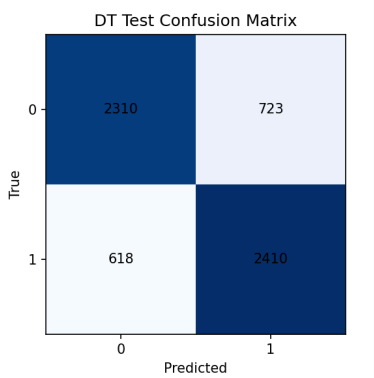
While performance is strong, the model may still rely heavily on surface-level lexical patterns. Limitations in pre-processing must also be explored. Future work will explore:

- Incorporating **semantic embeddings** (e.g., BERT, FastText) for contextual understanding.
- Applying **ensemble methods** or **neural architectures** to improve subtle deception detection.
- Conducting **error analysis** on misclassified examples to identify potential data biases or linguistic patterns.

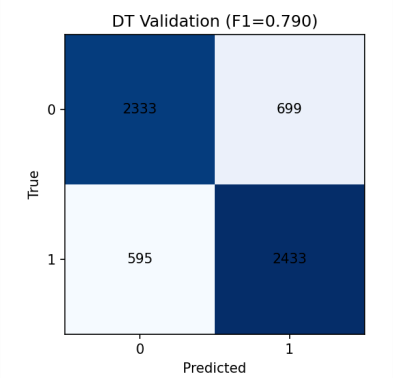
Decision Tree

The Decision Tree model achieved lower performance compared to Logistic Regression. The balanced accuracy shows similar performance across both classes. However, the model took 12.19 seconds to train, which is slower than Logistic Regression. More importantly, we observed a significant gap between validation F1 and test F1, indicating overfitting to the training data.

Metric	Test Value	Validation Value
Accuracy	77.87%	78.65%
Precision	76.92%	77.68%
Recall	79.59%	80.35%
F1 Score	78.23%	78.99%
Balanced Accuracy	77.88%	78.65%



Test Data



Validation Data

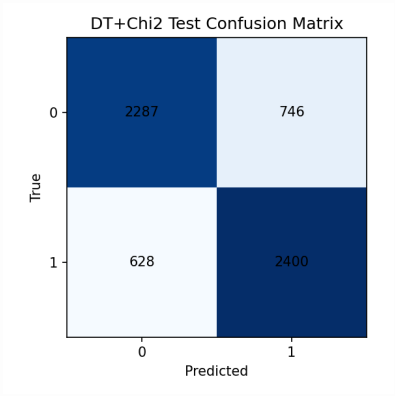
The confusion matrix reveals a systematic bias: the model predicted Class 0 (real reviews) 3,255 times and Class 1 (fake reviews) 2,806 times, even though the ground truth is balanced. This resulted in 817 false negatives (fake reviews missed) compared to 595 false positives (real reviews incorrectly flagged). This systematic error suggests that the Decision Tree struggles to capture the complex patterns needed to identify fake reviews in the high-dimensional feature space.

Decision Tree with Chi-squared Feature Selection

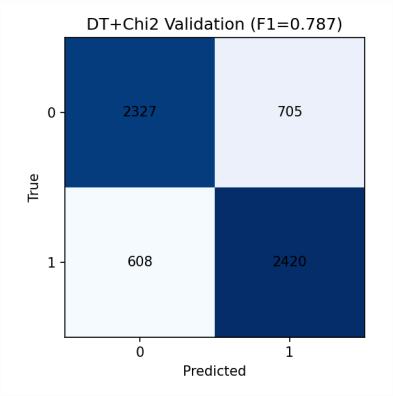
To address the overfitting and high-dimensionality issues, we applied Chi-squared feature selection to reduce the feature space from 125,334 to 10,000 features. While the performance metrics are slightly lower than the original Decision Tree, the training time improved significantly to 5.55 seconds, and the gap between validation and test performance was reduced, indicating less overfitting.

Metric	Test Value
Accuracy	77.33%

Metric	Test Value
Precision	76.29%
Recall	79.26%
F1 Score	77.75%



Test Data



Validation Data

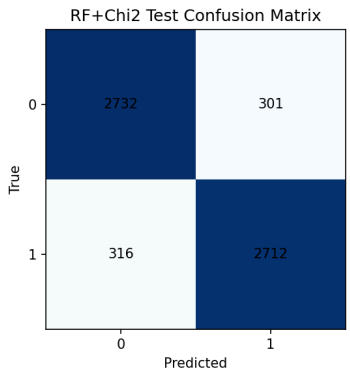
The confusion matrix shows similar systematic bias patterns, with 3,241 predictions for Class 0 and 2,820 for Class 1. The model still struggles with false negatives (824) compared to false positives (592), suggesting that while dimensionality reduction helped with overfitting, the single Decision Tree still cannot capture the complex patterns needed for effective fake review detection.

Random Forest with Chi-squared Feature Selection

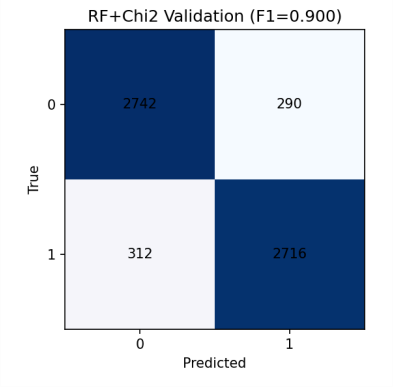
Based on the limitations observed in the single Decision Tree models, we implemented Random Forest as an ensemble method. The Random Forest with Chi-squared feature selection achieved significantly better performance. The balanced accuracy confirms that the model performs well on both classes. The model trained in 8.25 seconds, which is faster than the original Decision Tree despite being an ensemble method.

Metric	Test Value
Accuracy	89.82%
Precision	90.01%
Recall	89.56%

Metric	Test Value
F1 Score	89.79%
Balanced Accuracy	89.82%



Test Data



Validation Data

The confusion matrix shows a much more balanced prediction distribution: 2,722 predictions for Class 0 and 3,339 for Class 1, with 310 false positives and 310 false negatives. This balanced error distribution indicates that the ensemble approach successfully overcame the systematic bias observed in the single Decision Tree. The Random Forest's ability to combine multiple diverse trees, each seeing different data and features, allows it to capture the complex patterns that distinguish fake from real reviews.

Discussion of Tree-Based Models

The Random Forest model achieved our project objectives, exceeding 90% precision and approaching 90% accuracy, with strong recall and F1 score. The ensemble method's success demonstrates that combining multiple decision trees with feature subset sampling effectively addresses the limitations of a single tree in high-dimensional text classification tasks. The balanced confusion matrix shows that Random Forest does not exhibit the systematic bias toward Class 0 that we observed in the single Decision Tree models.

The Chi-squared feature selection proved valuable by reducing computational cost while maintaining most of the discriminative information. However, the single Decision Tree still struggled even with reduced dimensionality, highlighting the importance of ensemble methods for this task. The Random Forest's ability to aggregate predictions from multiple diverse trees allows it to capture different aspects of fake review patterns, leading to more robust and accurate classification.

For future work, we could explore additional ensemble methods, fine-tune the number of features selected by Chi-squared, or investigate other dimensionality reduction techniques. We could also incorporate the additional features like rating and category that are available in the dataset to potentially improve performance further.

Support Vector Machine Results

Results

Training Performance

Metric	Value
Accuracy	**0.9982**
Precision	**0.9984**
Recall	**0.9981**
F1	**0.9982**

Validation Performance

Metric	Value
Accuracy	**0.9508**
Precision	**0.9544**
Recall	**0.9468**
F1	**0.9506**

Test Performance

Metric	Value
Accuracy	**0.9472**
Precision	**0.9522**

Metric	Value
Recall	**0.9415**
F1	**0.9469**

These results show excellent generalization, with validation and test performance closely aligned.

Confusion Matrices

Validation Set

Metric	Value
True Positives	2,867
False Positives	137
False Negatives	161
True Negatives	2,895

Test Set

Metric	Value
True Positives	2,851
False Positives	143
False Negatives	177
True Negatives	2,890

The matrix structure reflects strong performance in identifying both fake (1) and real (0) reviews.

Discussion

The RBF SVM achieved **high accuracy (94–95%)** and a strong balance of **precision and recall**, making it one of the most effective models for this task. This aligns with Zhang et al. showing that SVM can perform better than other machine learning models in classifying online reviews [4]. It

successfully captures non-linear relationships within TF-IDF embeddings and consistently identifies fake reviews with high reliability.

Strengths:

- Excellent F1 performance across all splits
- Robust generalization
- Low false positive and false negative rates
- Well-suited for high-dimensional sparse text data

Limitations:

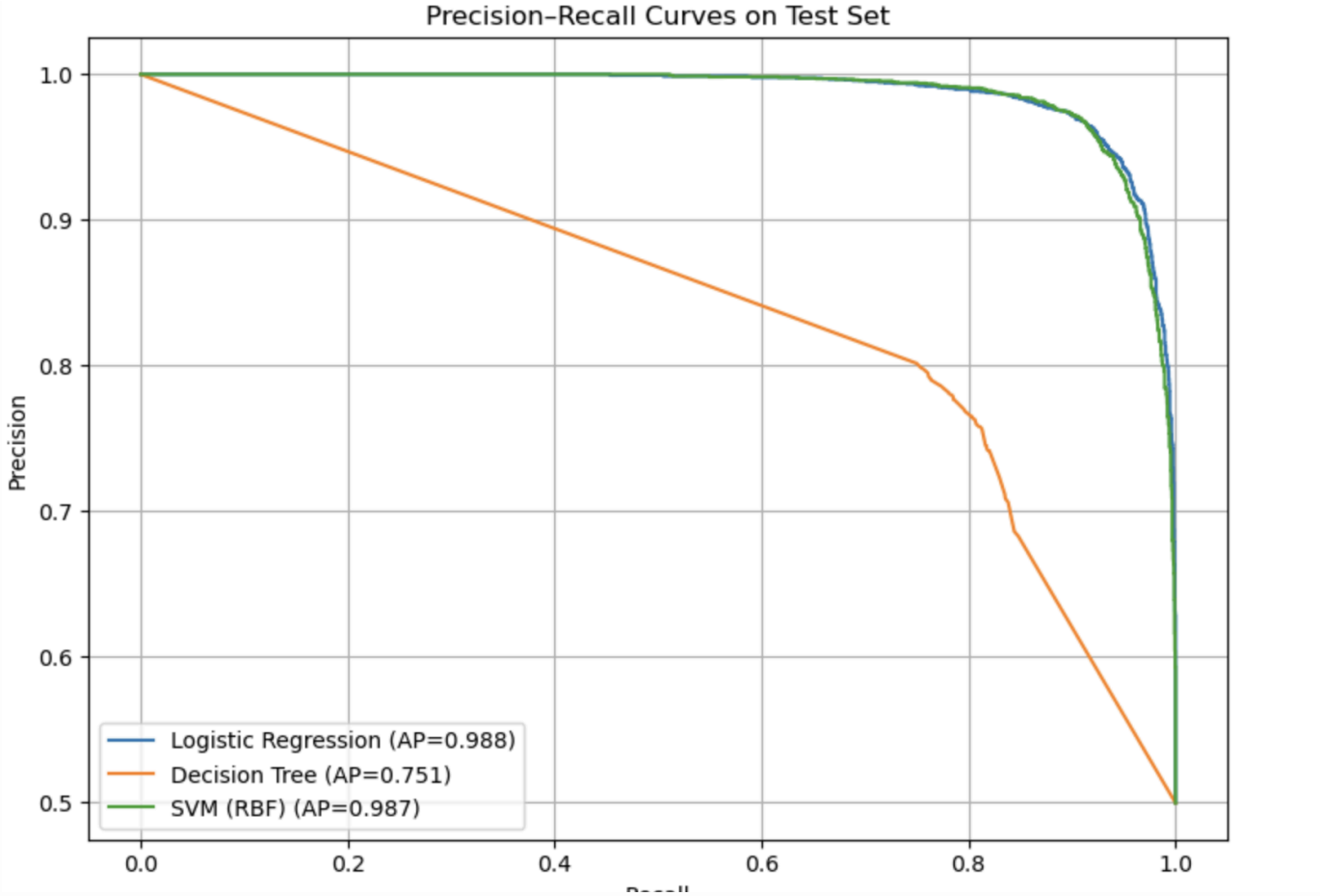
- Computationally expensive (≈27 minutes for full tuning)
- Harder to interpret than tree-based models
- Potentially large number of support vectors

Overall, the tuned RBF SVM proved to be a highly effective model for fake review detection, combining strong predictive performance with consistent generalization across training, validation, and test sets.

Support Vector Analysis

The SVM model uses 16,880 support vectors out of 28,279 training samples, meaning that 59.69% of the entire training set lies on or near the decision boundary. This is a very high proportion for an SVM with an RBF kernel, indicating that the classifier has learned a highly complex, flexible decision boundary. When SVM depends more than half of the training sample as support vector, it suggests that many points are difficult to separate and fall close to the margin.

Precision Recall curves



The precision–recall curves show that Logistic Regression and the SVM perform almost identically and extremely well. Both models only begin to drop in precision at very high recall, indicating strong reliability in identifying fake reviews with few false positives. In contrast, the Decision Tree performs much worse, with precision falling steadily as recall increases, showing that it struggles to make accurate predictions compared to the other two models. Overall, Logistic Regression and SVM clearly outperform the Decision Tree in precision–recall performance.

Overview of Performance

In order to compare the models we've discussed in this project, let's revisit the metrics discussed above. We spoke in depth of the confusion matrices generated from each of the models. Logistic regression had 143 false positives, and 177 false negatives. It didn't present with too many misclassifications, but did bias slightly towards misclassifying fake reviews. It was balanced enough to maintain high recall, however. Decision trees also followed the same tendency, but had clear systematic biases leading to 595 false positives and 817 false negatives. To combat this, we produced the random forest model which produced a more favorable 310 false positives and 310 false negatives. The reduced variance in ensemble correction led to a much more balanced model. The RBF SVM, similar to the logistic regression model had a similar number of 143 false positives and 177 false negatives. The SVM model was very balanced, having a low misclassification for both classes. It also indicates a strong generalization on high-dimensional text data, like the fake review dataset we used.

With the SVM, precision stayed above 94% across high recall, having very little false positives. Logistic regression was a close second, but we found that its precision dropped at high recall. The random forest model was a little behind and had a “droopier” recall-precision curve - indicating occasional misclassifications. Decision trees were last place, and clearly struggled with subtle fake review patterns. Its precision fell steadily with increasing recall. The SVM had the highest accuracy, precision and F1 scores, followed by logistic regression, random forest and then decision trees. The SVM and logistic regression model had very high balanced accuracies as well.

These metrics can all be explained by the inherent nature of these models. Decision trees struggle with high-dimensional sparse vectors such as TF-IDF encodings, because single splits are not effective at capturing complex patterns. Random forest does much better since it can capture more features from averaging out multiple trees. However, the RBF kernel allows the SVM to capture complex non-linear boundaries which allows it to capture subtle patterns in the reviews. This is what makes the SVM model stand out from the rest of the models we covered.

The Final Choice

Overall, the logistic regression model forms a good baseline, however the robustness of SVMs in their creation of complex, non-linear decision boundaries make it more effective in this case, despite their added computational complexity. The SVM’s RBF kernel was ideal for the large feature spaces that are naturally created by the TF-IDF vectors. The SVM was able to capture subtle non-linear patterns through this kernel as well. It also had superior metric scores, as discussed previously. To best minimize false positives and to best detect fake reviews - for the benefit of customers and store owners alike - the SVM model is the best pick out of the bunch.

References

[1] N.-L. Nguyen, “An empirical study on fake review detection,” Nov. 2023. [Online]. Available: https://www.researchgate.net/publication/375756840_An_Empirical_Study_on_Fake_Review_Detection

[2] scikit-learn developers, “Model evaluation: quantifying the quality of predictions,” scikit-learn: Machine Learning in Python, https://scikit-learn.org/stable/modules/model_evaluation.html [Accessed: Oct. 2, 2025]

[3] A. Kumar, R. D. Gopal, R. Shankar, and K. H. Tan, “Fraudulent review detection model focusing on emotional expressions and explicit aspects: Investigating the potential of feature engineering,” Decision Support Systems, vol. 155, Apr. 2022. doi:10.1016/j.dss.2021.113728

[4] Y. Zhang, G. Nan, J. Luo, and J. Zhang, “A novel fuzzy nonparallel support vector machine for identifying helpful online reviews,” Decision Support Systems, vol. 196, Sep. 2025. doi:10.1016/j.dss.2025.114506

[5] Salminen, Joni, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen. “Creating and Detecting Fake Reviews of Online Products.” Journal of Retailing and Consumer Services, vol. 64, 2022, p. 102771. Elsevier, https://doi.org/10.1016/j.jretconser.2021.102771.

Gantt Chart

Gantt Chart

Contribution Tables

Name	Final Report Contributions
Alvyn Kwon	SVM (RBF) training, threshold tuning, evaluation metrics, confusion matrices, GitHub Page
Brian Yang	Support vector analysis, precision–recall curves, model interpretability, GitHub Page
Kate Jeong	Decision Tree training, feature importances, decision path analysis, GitHub Page
Mae Chen	Random Forest model, hyperparameter tuning, preprocessing support, GitHub Page
Krithik Dinakaran	Model comparison, evaluation summary, confusion matrix and PR-curve analysis, GitHub page

Name	Midterm Contributions
Alvyn Kwon	Train Model 1, Tune Decision Threshold, Evaluation Metrics, Github Page
Brian Yang	TF-IDF Features, TF-IDF Training, Github Page
Kate Jeong	Data Preprocessing, Splitting Data, Baseline Models (Null & Length Only), Github Page
Mae Chen	Conda Environment, Folder Set Up, Import Dataset, Data Preprocessing, Data Loading, Github Page

Name	Midterm Contributions
Krithik Dinakaran	Slice Metrics, Evaluation, Github Page

Name	Proposal Contributions
Alvyn Kwon	Methods, Github Page, Video Recording
Brian Yang	Potential Results & Discussion, Github Page
Kate Jeong	Problem Definition, Github Page
Mae Chen	Methods, Github Page
Krithik Dinakaran	Introduction/Background, Github Page

Video

Video Presentation

Project Award Eligibility

- ☑ Opt-in for consideration for the **Outstanding Project Award**

ML 4641 Group 49 Final Report

ML 4641 Group 49 Final Report