Fake Reviews Prediction

Ninad Shah
MS, Data Science
Tempe, AZ
nshah103@asu.edu

Ashray Kothari
MS, Computer Science
Tempe, AZ
akotha14@asu.edu

Khushkumar Kantaria
MS, Computer Science
Tempe, AZ
kkantari@asu.edu

Kavya Parikh
MS, Computer Science
Tempe, AZ
kparik10@asu.edu

Kaumudi Patil
MS, RAS-AI
Tempe, AZ
kspatil1@asu.edu

***Abstract*—Consumer trust and decision-making processes are impacted by the prevalence of fraudulent reviews, which seriously undermines the legitimacy and dependability of online review sites. In this work, we investigate machine learning and deep learning algorithms to tackle this urgent problem on two major platforms: Yelp and Amazon.**

Utilizing an extensive comparison study, we investigate multiple approaches that include feature engineering, data preprocessing, and model selection for efficiently identifying genuine and fake reviews. Our tests indicate encouraging outcomes, demonstrating the effectiveness of our suggested methods in precisely identifying fraudulent reviews.

Additionally, we explore the field of fake review detection systems, exploring diverse methods from both traditional statistical machine learning and deep learning. We critically evaluate the efficacy of current datasets, data-gathering procedures, and feature extraction approaches.

Furthermore, we perform benchmark tests on benchmark datasets from Yelp and Amazon using neural network models and transformers, highlighting the need for improved review filtering and categorization techniques. Our results highlight how important it is to strengthen credibility and trust in e-commerce platforms by implementing better review authentication systems, especially in light of the growing problems caused by bogus reviews.

***Index Terms*—Fake review detection, Machine learning algorithms, Hyperparameter tuning, Amazon dataset, Yelp dataset, Test accuracy, F1 score, Model evaluation**

I. Introduction

A. Background and Problem Statement

The pervasive presence of fake reviews on prominent e-commerce platforms such as Amazon and Yelp undermines the credibility of online feedback systems, erodes consumer trust, and poses significant challenges to businesses and consumers alike. Despite concerted efforts to combat this issue, the proliferation of deceptive reviews persists, necessitating innovative and effective detection mechanisms to safeguard the integrity of online platforms.

Within the realm of fake review detection, the datasets sourced from Amazon and Yelp [3] serve as crucial battlegrounds for identifying fraudulent activities and preserving the authenticity of user-generated content. However, existing approaches often face limitations in accurately discerning between genuine and fabricated reviews within these datasets, thereby impeding the efficacy of detection efforts. The challenge lies in developing robust methodologies capable of effectively distinguishing between authentic and fake reviews amidst the vast volume of user-generated content on Amazon and Yelp. Traditional statistical machine-learning techniques and more advanced deep-learning methods have been explored for this purpose, yet there remains a pressing need for comprehensive solutions that address the evolving tactics employed by malicious actors.

Fig. 1. Yelp: Word cloud for Actual reviews

Furthermore, the lack of standardized datasets and feature extraction techniques specific to Amazon and Yelp datasets complicates the development of accurate and scalable detection models. To bridge this gap, there is a demand for in-depth research that comprehensively analyzes the nuances of these datasets, identifies key features indicative of fake reviews, and evaluates the performance of detection techniques across diverse scenarios.

Fig. 2. Amazon: Word cloud for reviews

In light of these challenges, this research endeavors to conduct a thorough investigation into fake review detection on Amazon and Yelp datasets. By leveraging a combination of traditional statistical machine learning algorithms and cutting-edge deep learning methodologies, the study aims to enhance

the accuracy and efficiency of detection models tailored to these platforms.

Through the analysis of existing datasets, exploration of feature extraction techniques, and critical evaluation of detection methodologies, this research seeks to contribute to the development of robust frameworks for detecting fake reviews on Amazon and Yelp. By establishing benchmarks and identifying best practices, the study aims to empower businesses and consumers with reliable tools for discerning authentic feedback from deceptive content, thereby fortifying the credibility of online platforms and fostering trust within the e-commerce ecosystem.

B. Objective

The project aims to create a fake review detection system using sentiment analysis, deep neural networks, and NLP. It focuses on identifying patterns to distinguish genuine user reviews from potentially manipulated/fabricated ones generated by bots, paid reviewers, or individuals with malicious intent, addressing challenges like misleading consumers, harming platform reputation, and ethical/legal concerns.

C. Importance

Predicting fake reviews is a very important endeavor for protecting several aspects of the digital ecosystem. Fundamentally, the goal of this project is to safeguard consumer confidence and trust by guaranteeing that people can depend on reliable reviews when making judgments about what to buy. Furthermore, by identifying and addressing fraudulent reviews, the initiative preserves the integrity of online platforms and promotes a culture of trust and honesty among users. Since legitimate companies have an even playing field on which to display their goods and services, fair competition is also maintained.

Furthermore, by optimizing the review validation process and empowering platforms to manage massive amounts of data, the project increases scalability and efficiency. In addition to these basic justifications, handling fake reviews is essential for maintaining brand reputation, legal and regulatory compliance, user experience, stopping fraudulent activity, helping small businesses, and, in the end, encouraging accountability and transparency in the digital marketplace.

D. Existing Literature

This study employs machine learning techniques to address the pervasive issue of fraudulent reviews in e-commerce, aiming to filter and categorize reviews effectively. By reviewing prior research in sentiment analysis, natural language processing, and e-commerce review analysis, the project seeks to develop robust methods for identifying fake reviews and assessing review credibility.

Through a comprehensive exploration of grouping and classification algorithms, including supervised and unsupervised learning approaches, the study underscores the importance of leveraging machine learning to combat fraudulent reviews and enhance the integrity of online platforms, ultimately aiming to improve the online shopping experience for customers, despite challenges such as limited dataset availability.

1) For Yelp Dataset: The literature review provides a comprehensive overview of research on identifying fraudulent reviews in Yelp's dataset, emphasizing their impact on consumer decisions. It examines four prominent machine learning classification techniques: XGBoost, Support Vector Machine, Gaussian Naïve Bayes, and Logistic Regression, detailing their principles and suitability for the task. The methods section describes the dataset, feature engineering, and data preprocessing, addressing challenges like unbalanced data and identifying key characteristics.

Statistical evaluations reveal insights such as the tendency for fraudulent reviews to have extreme ratings. Evaluation metrics like the F-1 score demonstrate XGBoost's effectiveness, outperforming other techniques with a score of 0.99. The study concludes with a summary of results, acknowledging dataset limitations, and proposing enhancements like incorporating user trust factors to advance fake review detection on platforms like Yelp [1].

2) For Amazon Dataset: Despite extensive research over two decades, fraudulent reviews persist, impacting consumer decisions and market outcomes. Our study addresses this challenge using a specialized dataset tracking products that purchase fake reviews. Our analysis reveals a notable trend: products engaging in dishonest conduct are clustered within the reviewer network, relying on common reviewers rather than genuine feedback providers. Leveraging this network structure, we introduce a novel method surpassing traditional text analysis techniques in accurately identifying fraudulent review buyers.

Our approach offers resilience to seller manipulation, adaptability across different environments, and transferability to various platforms beyond Amazon. By focusing on network properties instead of review text, our method provides a more effective and efficient means of detecting fraudulent behavior, with potential implications for enhancing trust in online review systems. Furthermore, our findings highlight the importance of targeted measures to deter review manipulation by uncovering its financial motivations. Beyond aiding platforms in combating fraud, our research contributes to a deeper understanding of the dynamics of the online market [2].

3) Sentimental Analysis: We looked at the inner workings of Yelp's phony review detection system in a previous study, which concentrated on supervised learning with training from Yelp's filtered reviews. They examined actual fraudulent reviews that Yelp had detected, in contrast to earlier methods that relied on pseudo-fake reviews. Based on actual Yelp data, their analysis demonstrated the superiority of behavioral traits over linguistic ones.

To identify the psycholinguistic differences between commercial reviews screened by Yelp and crowdsourced phony reviews, they put forth a novel information-theoretic analysis. Through testing, they were able to obtain a high classification accuracy, confirming that Yelp's filtering algorithm is good at

identifying unusual spamming activities. These findings are relevant to other systems that may be in place.

Prior research investigated sentiment analysis [4] techniques, specifically for text-based tweets, with the goal of classifying them as neutral, negative, or positive. Term frequency-inverse document frequency (TF-IDF) and word embedding were the two main word representation approaches they examined, and they examined classification metrics on eight different datasets. Because word embeddings are better at contextual understanding than classic TF-IDF approaches, their investigation showed that Recurrent Neural Network (RNN) models containing word embeddings regularly outperformed other methods.

4) BERT: Bidirectional Encoder Representations from Transformers, or BERT [5], pre-trains deep bidirectional representations from unlabeled text, taking into account both left and right context in every layer, so revolutionizing natural language understanding. This technique improves the adaptability and performance of the model across a range of activities by enabling fine-tuning for certain downstream tasks without requiring large architectural alterations.

E. System Overview

- In the data collection part, we compiled a varied dataset from sources, such as Amazon reviews dataset and Yelp businesses information dataset, that includes both real and fraudulent reviews.
- In the data pre-processing step, we used techniques for imputation or removal to handle any missing data or punctuation; broke down reviews into individual words or phrases; and converted text data into numerical features using techniques like TF-IDF (Term Frequency-Inverse Document Frequency).
- We conducted an in-depth analysis of the information present in the dataset to understand the relationship between fake vs. genuine reviews and also visualized key patterns and trends in the data.
- As a part of the feature engineering step in the pipeline, we attempted to understand the user behavior patterns to enhance the ability of the model to make accurate predictions.
- For model selection, throughout the project we focus on different models focusing on machine learning, deep learning, and natural language processing for achieving high accuracies.
- The dataset is split into training and validation sets selected models are trained on the training set and models are evaluated.
- When we talk about the model evaluation metrics, what we have particularly focused on is accuracy, precision, recall, and F-1 score.
- To increase the efficiency and improve the performance of the models, hyper-parameter tuning is performed.
- Deep analysis and feature engineering can be completed before we incorporate neural network models into the model selection procedure. Put into practice neural network designs that are appropriate for tasks involving natural language processing, such as BERT.
- Make use of visualization approaches to provide the models' performance metrics understandably and straightforwardly. For a thorough understanding of each model's performance on the validation set, plot the metrics for accuracy, precision, recall, and F-1 score. To see how these metrics change for various models or hyperparameter setups, use line plots or bar charts. To further understand how well the model performs in classification—particularly in identifying authentic from counterfeit reviews—visualize the confusion matrix.

II. Important Definitions

A. Data

Data refers to the information that is collected from e-commerce websites, social media platforms, etc. In this project, we are focussing on Amazon [6] and Yelp [7] datasets. Amazon dataset focuses on the users' reviews for different products while the Yelp dataset focuses on information about businesses across 8 metropolitan areas in the USA and Canada.

B. Prediction Target

The prediction target remains the same, which is to classify reviews as genuine or fake based on the content and metadata. This is a binary classification task where the model predicts whether a review is legitimate or created with malicious intent.

C. Variables or concepts in the data

- Textual Data: Amazon and Yelp reviews containing the text data
- Metadata: Content Text, category, ratings, sentiment, and word count category
- Sentiment Analysis: Labels indicating the sentiments expressed in the reviews - positive (1 if the rating is greater than or equal to 3) and negative (0 if the rating is less than 3).
- Linguistic features: Patterns or characteristics in the text that can be extracted, such as punctuation usage, frequency of certain words or phrases, grammatical structures, etc.
- User Behavior: Patterns related to the user activity

D. Constraints

- Platform-specific features / Feature discrepancy: It was difficult to manage the variety of features on many platforms. And, thus we had to make different models focussed on different datasets.
- Unbalanced dataset: Handling the class imbalance, in which the proportion of real reviews surpassed that of fraudulent reviews (or vice versa), needed caution. We investigated methods such as class weighting, resampling, and sophisticated algorithms built for unbalanced data to enhance the model's precision in identifying real and fraudulent reviews.

- Model accuracy and efficiency: To detect fake reviews, it was essential to strike a balance between computational efficiency and high accuracy. The appropriate balance between accuracy and efficiency was attained by model architecture optimization, hyperparameter optimization, and the use of strategies like feature reduction and model pruning.
- Scalability: Scalable solutions were required to handle enormous datasets efficiently to identify false reviews across numerous platforms. For efficient resource management and seamless model training and deployment, it was crucial to optimize algorithms, data pipelines, and distributed computing frameworks.

E. Algorithm

The fields of machine learning and unsupervised learning are combined. Supervised learning uses algorithms to learn from pairs of labeled data in order to provide predictions or classifications, whereas unsupervised learning uses algorithms to find patterns or structures in unlabeled data. Clustering is an unsupervised learning technique that groups data items according to comparable attributes. Sentiment analysis, which ascertains the sentiment conveyed in text, is one aspect of Natural Language Processing (NLP), which is concerned with how computers interact with human language. In order to improve the performance of machine learning models, feature engineering entails choosing or producing pertinent features from raw data. In machine learning applications, algorithms like Random Forest and Logistic Regression are frequently employed. Random Forest uses numerous decision trees to increase accuracy, while Logistic Regression estimates the probability of a binary result depending on input data. In many different fields, these algorithms are essential for data analysis and prediction/classification.

III. Analysis of the Given Dataset

There are two primary components to the suggested method/system: one for the Yelp dataset and another for the Amazon dataset. Using a variety of text preprocessing techniques and feature engineering methodologies, a broad range of machine learning models and classifiers, including SVMs, Naive Bayes, decision trees, and ensemble methods, were assessed for the Amazon dataset. A thorough comparison of each model's efficacy in detecting false reviews was produced by evaluating its performance using criteria including accuracy, precision, recall, and F1 score. In contrast, the Yelp dataset underwent data analysis that concentrated on pretreatment measures such as readability score computation, duplication removal, and sentiment analysis using the VADER sentiment analyzer. The links between the numerical characteristics were found using correlation analysis, and the most important features for the identification of phony reviews were discovered by feature importance analysis with a RandomForestRegressor.

The data analysis as a whole demonstrated how crucial feature engineering and preprocessing stages are to maximizing the effectiveness of fake review detection systems. Furthermore, the assessment of diverse machine learning models yielded significant insights into their relative efficacy across diverse datasets and scenarios.

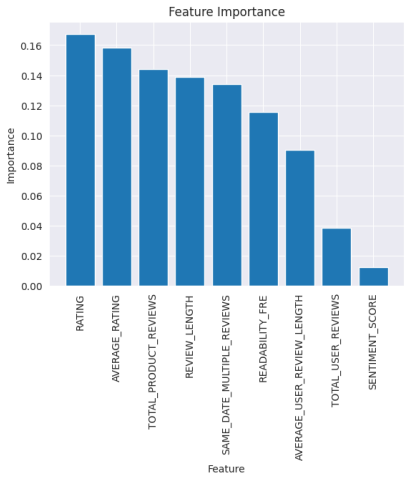

Fig. 3. Yelp: Feature Importance

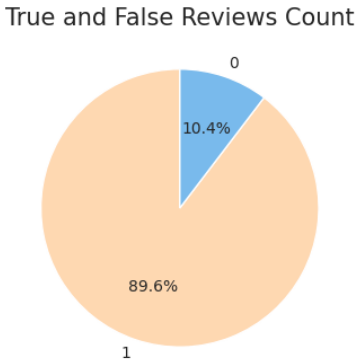

Fig. 4. Yelp: True and False Reviews Count

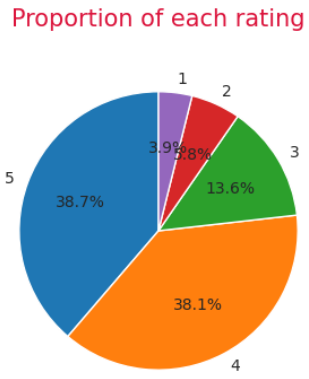

Fig. 5. Yelp: Proportion of each rating

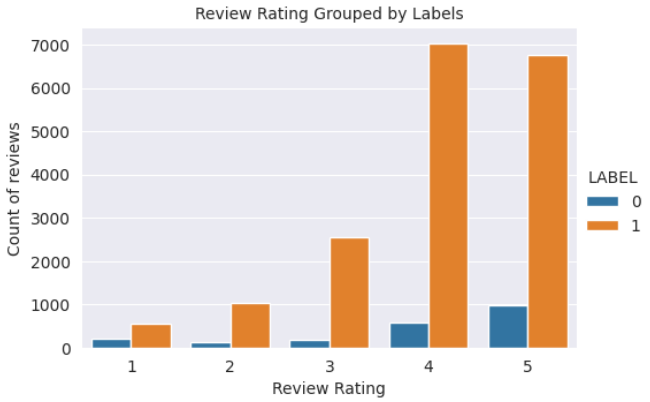

Fig. 6. Yelp: Review Rating Grouped by Labels

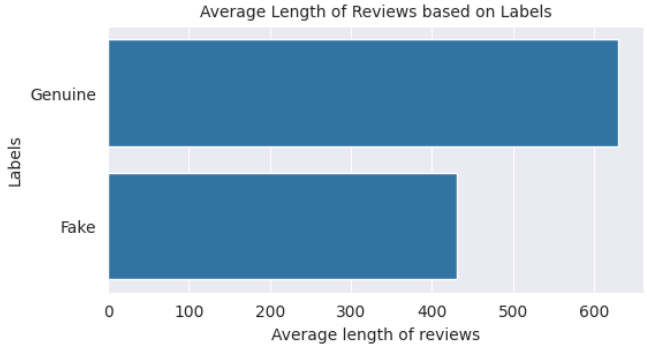

Fig. 7. Yelp: Average lengths of reviews

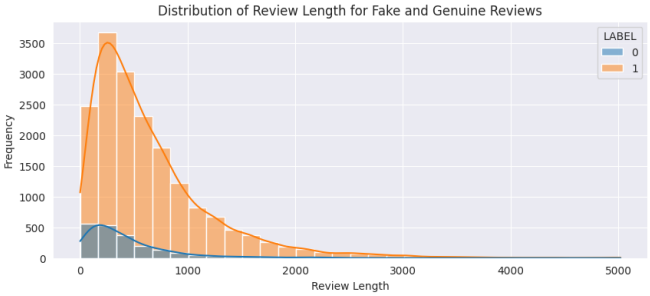

Fig. 8. Yelp: Distribution of review lengths

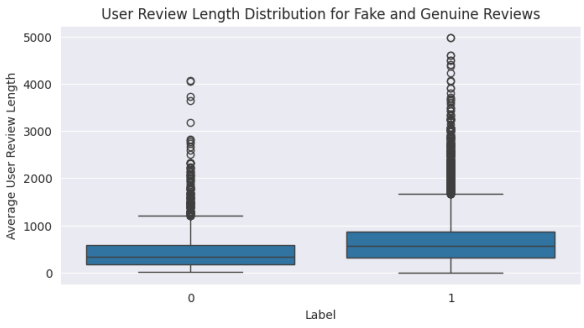

Fig. 9. Yelp: Distribution of User review length

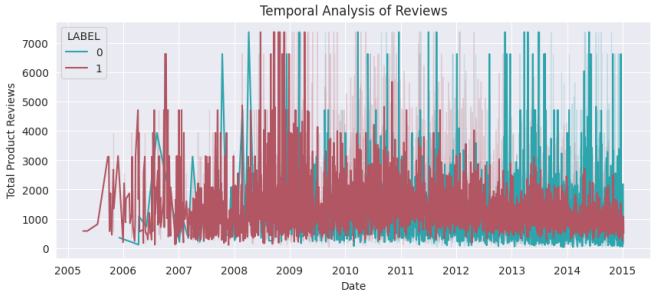

Fig. 10. Yelp: Temporal analysis of reviews

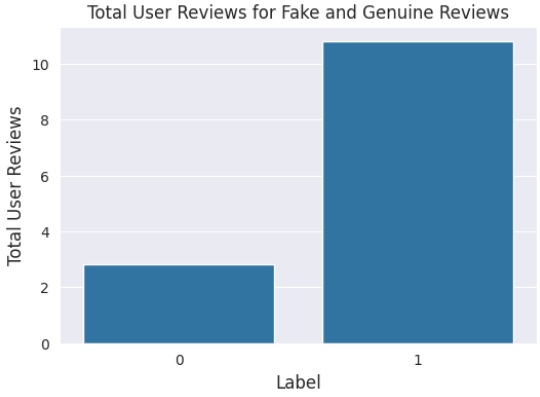

Fig. 11. Yelp: Total user reviews

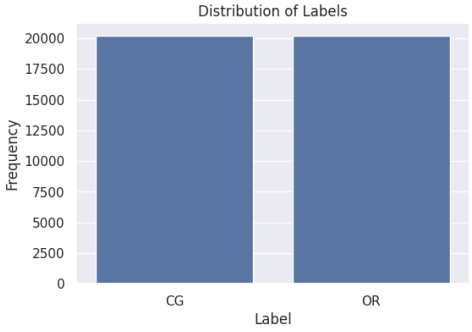

Fig. 12. Amazon: Distribution of labels

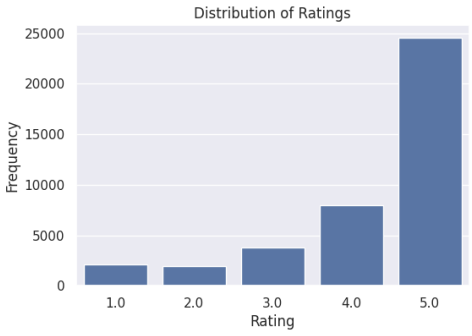

Fig. 13. Amazon: Distribution of Ratings

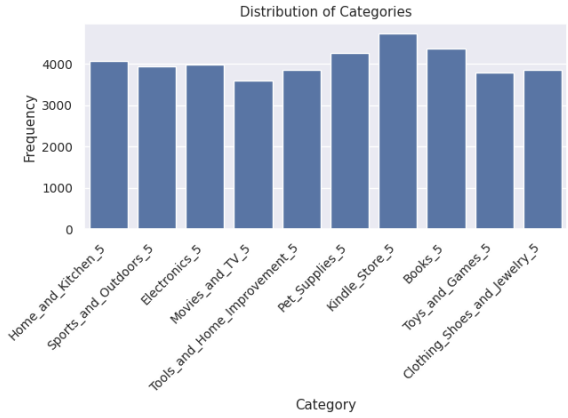

Fig. 14. Amazon: Distribution of Categories

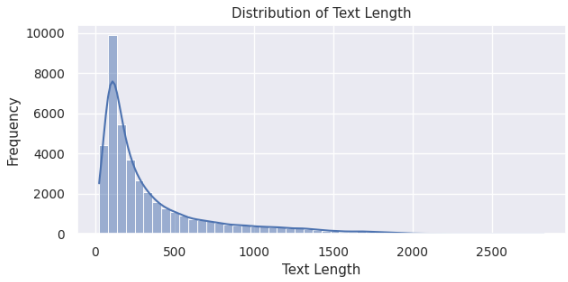

Fig. 15. Amazon: Distribution of Text Length

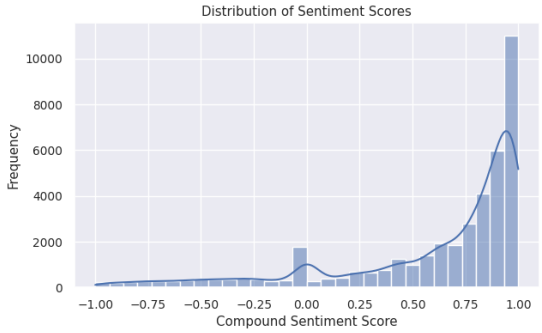

Fig. 16. Amazon: Distribution of sentiment score

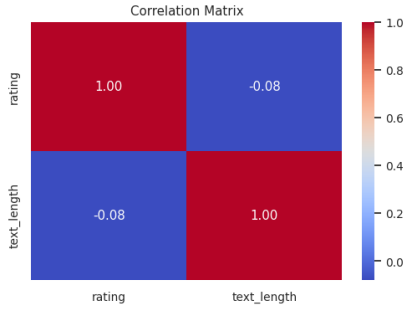

Fig. 17. Amazon: Correlation Matrix

IV. Traditional Methodologies and its Results

A. For Yelp Dataset

For the Yelp dataset, we performed data preprocessing tasks, including data importation, sentiment analysis using NLTK's Vader, readability analysis using the Flesch Reading Ease score, and data exploration and cleaning. Subsequently, we delved into machine learning classification tasks, where we imported classifiers from scikit-learn, preprocessed text data, split the dataset, performed feature scaling, trained classifiers, and evaluated their performance using metrics like accuracy, precision, and F1-score. The code also conducted thorough evaluations for each classifier, generating confusion matrices, ROC curves, precision-recall curves, and calibration curves to gain insights into classifier performance across various metrics.

TABLE I
Model Performance Metrics - Accuracy, Recall, and F1-Score

Model	Accuracy	Recall	F1-Score
LinearSVC	0.89700	1.000000	0.945704
MultinomialNB	0.89700	1.000000	0.945704
RidgeClassifier	0.89700	1.000000	0.945704
SGDClassifier	0.89700	1.000000	0.945704
BernoulliNB	0.89700	1.000000	0.945704
LogisticRegression	0.89700	1.000000	0.945704
SVM	0.89700	1.000000	0.945704
RandomForestClassifier	0.89650	0.994147	0.945151
LGBMClassifier	0.89575	0.994705	0.944805
AdaBoostClassifier	0.89575	0.994147	0.944776
CatBoostClassifier	0.89150	0.991918	0.942532
XGBClassifier	0.88600	0.980769	0.939151
KNeighborsClassifier	0.88400	0.976589	0.937901
BaggingClassifier	0.87350	0.960145	0.931585
ExtraTreeClassifier	0.82900	0.900502	0.904282
DecisionTreeClassifier	0.82225	0.889632	0.899789

TABLE II
Model Performance Metrics - Precision and ROC-AUC

Model	Precision	ROC-AUC
LinearSVC	0.897000	None
MultinomialNB	0.897000	0.701393
RidgeClassifier	0.897000	None
SGDClassifier	0.897000	None
BernoulliNB	0.897000	0.719625
LogisticRegression	0.897000	0.736405
SVM	0.897000	None
RandomForestClassifier	0.900758	0.763714
LGBMClassifier	0.899672	0.764374
AdaBoostClassifier	0.900076	0.764949
CatBoostClassifier	0.897830	0.770613
XGBClassifier	0.900922	0.748119
KNeighborsClassifier	0.902163	0.615842
BaggingClassifier	0.904674	0.713847
ExtraTreeClassifier	0.908094	0.553406
DecisionTreeClassifier	0.910180	0.562535

The same performance measures, including accuracy, precision, and an F1 score of 89.70%, were attained by a variety of machine learning models, including RidgeClassifier, SGDClassifier, BernoulliNB, LogisticRegression, and LinearSVC. Strong overall performance was shown by these models,

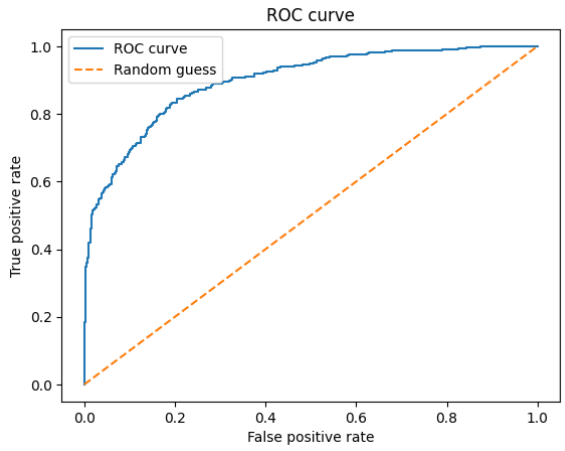

Fig. 18. Yelp: ROC curve for Logistic Regression

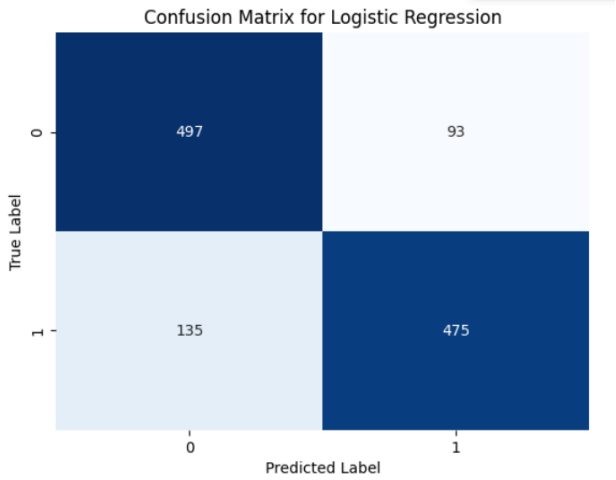

Fig. 19. Yelp: Confusion matrix for Logistic Regression

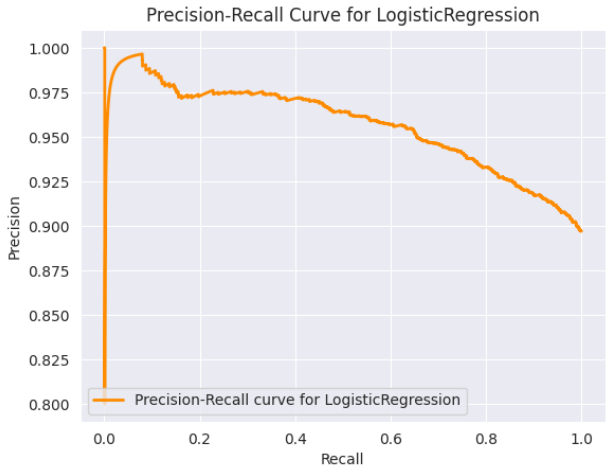

Fig. 20. Yelp: Precision-Recall for Logistic Regression

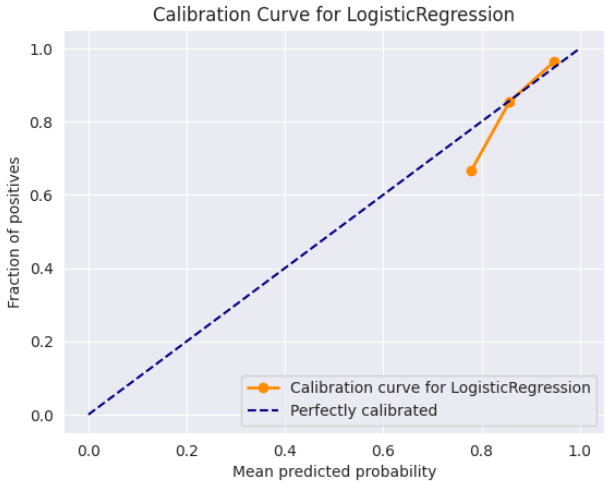

Fig. 21. Yelp: Calibration curve for Logistic Regression

indicating that they were successful in differentiating between the dataset's classes. With accuracy scores of 89.57% and precision scores of 89.97% and 90.01%, respectively, LGBMClassifier and AdaBoostClassifier closely trailed, demonstrating their resilience in classifying tasks.

Additionally performing well were RandomForestClassifier and CatBoostClassifier, with accuracy scores of 89.38% and 89.15%, respectively. However, models with accuracy ratings ranging from 81.65% to 88.40% showed somewhat poorer performance, such as KNeighborsClassifier, BaggingClassifier, DecisionTreeClassifier, and ExtraTreeClassifier. Based on the available metrics, LinearSVC, MultinomialNB, RidgeClassifier, SGDClassifier, BernoulliNB, and LogisticRegression were shown to be the best-performing models overall.

B. For Amazon dataset

For the amazon dataset, we implemented various machine learning models, including Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and others, to classify text data effectively. The models were trained and evaluated on a dataset comprising features like category, rating, and text content, facilitating a comprehensive comparison of their performance.

Performance metrics such as accuracy, precision, recall, and F1 score were calculated for each model, providing insights into their effectiveness in text classification. By leveraging multiple models and visualizing metrics such as ROC curves, precision-recall curves, calibration curves, and confusion matrices, we gained valuable insights into the strengths and weaknesses of different algorithms for text classification tasks.

With an accuracy of 90.21%, Logistic Regression was the most accurate machine learning model among those assessed on the dataset; the SGD Classifier came in second with an accuracy of 89.85%. The models demonstrated strong performance in differentiating between authentic and fraudulent reviews, as seen by their elevated precision, recall, F1 Score, and ROC AUC values.

TABLE III
MODEL PERFORMANCE METRICS (ACCURACY, RECALL, F1 SCORE)

Model	Accuracy	Recall	F1 Score
Multinomial Naive Bayes	0.858786	0.858786	0.858467
Bernoulli Naive Bayes	0.726969	0.726969	0.713859
LinearSVC	0.798937	0.798937	0.798245
Random Forest	0.875850	0.875850	0.875626
Extra Trees	0.893533	0.893533	0.893397
AdaBoost	0.791270	0.791270	0.791264
Bagging	0.814146	0.814146	0.813977
Decision Tree	0.760356	0.760356	0.760323
Extra Tree	0.691975	0.691975	0.690990
K Nearest Neighbors	0.616050	0.616050	0.576590
Logistic Regression	0.902065	0.902065	0.902065
Ridge Classifier	0.817856	0.817856	0.816748
SGD Classifier	0.898479	0.898479	0.898405
CatBoost	0.890936	0.890936	0.890937
LightGBM	0.884630	0.884630	0.884605

TABLE IV
MODEL PERFORMANCE METRICS (PRECISION, ROC AUC)

Model	Precision	ROC AUC
Multinomial Naive Bayes	0.862643	0.944864
Bernoulli Naive Bayes	0.780629	0.887355
LinearSVC	0.803781	None
Random Forest	0.879133	0.955719
Extra Trees	0.896021	0.965717
AdaBoost	0.791412	0.885818
Bagging	0.815669	0.892338
Decision Tree	0.760660	0.760451
Extra Tree	0.695027	0.692368
K Nearest Neighbors	0.689391	0.748377
Logistic Regression	0.902151	0.967899
Ridge Classifier	0.826662	None
SGD Classifier	0.899357	None
CatBoost	0.890942	0.961765
LightGBM	0.884824	0.958063

CatBoost and Extra Trees both performed admirably, achieving 89.35% and 89.09% accuracy, respectively. However, K Nearest Neighbors had the lowest accuracy of any method, at 61.61%, indicating that it is not very useful in this situation. The best-performing model overall was Logistic Regression, demonstrating its applicability to tasks involving the detection of false reviews on this dataset.

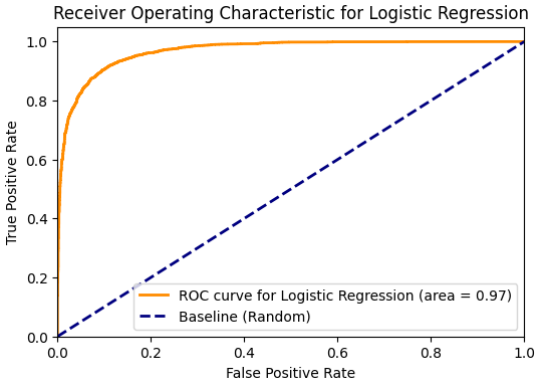

Fig. 22. Amazon: ROC for Logistic Regression

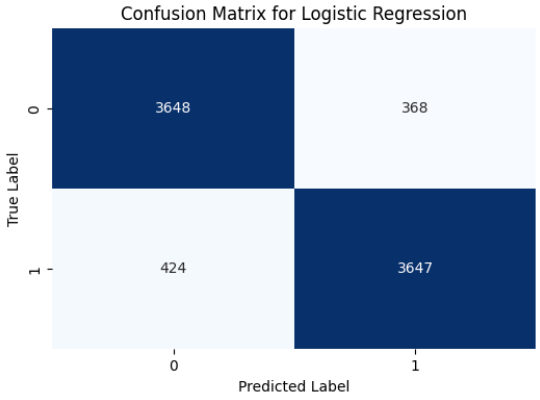

Fig. 23. Amazon: Confusion matrix for Logistic Regression

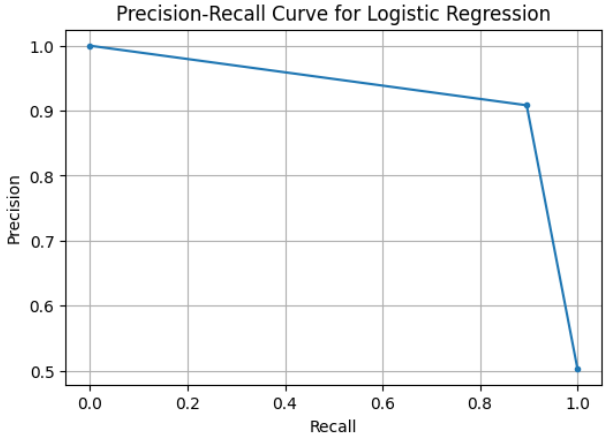

Fig. 24. Amazon: Precision-Recall for Logistic Regression

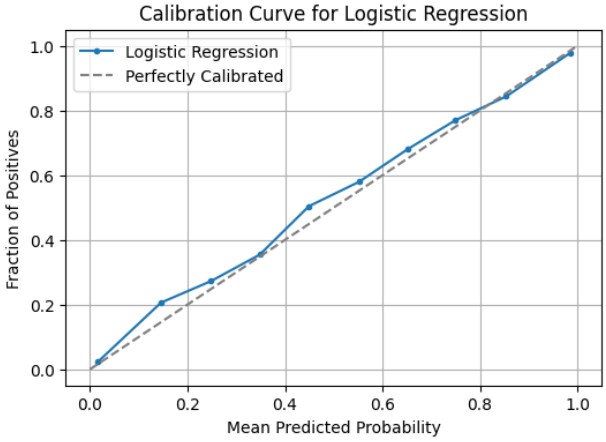

Fig. 25. Amazon: Calibration curve for Logistic Regression

V. APPROACHED METHODOLOGIES AND ITS RESULTS

A. For Yelp dataset

1) Neural Network Model:

- Trained a neural network model with a learning rate of 0.001 and 15 epochs.

- Used backpropagation to update the model's weights and biases during training.
- Monitored the training process to prevent underfitting or overfitting.
- For this approach, we obtained an accuracy of 83.04%.

2) Hyperparameter Tuning:

- Conducted hyperparameter tuning for the neural network model with a varying learning rate and epochs.
- We obtained the best results for a learning rate of 0.001 and 10 epochs.
- Utilized techniques like grid search or random search to find the optimal combination of hyperparameters that maximizes the model's performance.

3) Handling unbalanced datasets:

- Generated New Fake Reviews Using Data Augmentation Technique - Back Translation
  - Utilized the back translation technique for data augmentation, where original English text was translated into another language (e.g., French) using MarianMT and then translated back to English.
  - Retained the back-translated English text if it differed from the original English sentence, thereby generating new fake reviews.
- Generated Original Data Upsampled and Original Data Downsampled Datasets
  - Addressed label imbalance by creating two new datasets: original_data_upsampled and original_data_downsampled for the Yelp dataset.
  - Upsampled the minority class and downsampled the majority class to mitigate the issue of label imbalance.
- Determined Model Performance on Different Datasets
  - Observed that traditional models performed better on the new_data generated using data augmentation techniques compared to the original_data_upsampled and original_data_downsampled datasets.
  - Among the original_data, the upsampled version yielded higher performance metrics than the downsampled version.
- Evaluation Metrics and Visualization
  - Calculated accuracy, recall, precision, and F1 scores to evaluate model performance.
  - Plotted ROC curves to visualize the trade-off between true positive rate and false positive rate.

TABLE V
PERFORMANCE METRICS - NEW DATA
ACCURACY, RECALL, AND F1 SCORE

Model	Accuracy	Recall	F1 Score
Logistic Regression	0.810	0.779	0.806
SVC	0.773	0.715	0.762
Multinomial NB	0.782	0.693	0.764
Random Forest	0.745	0.692	0.734

- Features Used and Model Effectiveness
  - Included features such as text_final, rating, sentiment, and word_count_category for model training.
  - Word_count_category represents different bins based on the word counts of each review.
  - Found that the BERT model did not significantly improve performance for the new_data generated for Yelp.
- Deep Learning Model for New Data
  - Tuned hyperparameters for the deep learning model on the new_data.
  - Best hyperparameters obtained were 'learning_rate': 0.0001, 'num_epochs': 10, resulting in an accuracy of 83.04%.
- Original Data Upsampled - Deep Learning Model
  - Achieved an accuracy of 94.83% on the original_data_upsampled dataset using the deep learning model.

TABLE VI
PERFORMANCE METRICS - NEW DATA
PRECISION AND ROC-AUC SCORE

Model	Precision	ROC-AUC Score
Logistic Regression	0.836	0.897
SVC	0.816	0.850
Multinomial NB	0.851	0.866
Random Forest	0.781	0.824

B. For Amazon dataset

1) Methodology for Normal Neural Network:

- Utilized traditional feedforward neural networks for classification tasks.
- Inputted features include text data, category, and rating.
- Experimented with different architectures, activation functions, and optimization algorithms.
- Trained the model using backpropagation and update weights through gradient descent.
- Evaluated performance using metrics such as accuracy, precision, recall, and F1 score.

2) Hyperparameter Tuning for Neural Networks:

- Explored various hyperparameters like learning rate, batch size, number of layers, and neurons per layer.
- Used techniques such as grid search and random search to find optimal hyperparameters.
- Employed cross-validation to assess model generalization and prevent overfitting.
- Regularized the network with techniques like dropout or L2 regularization to improve performance.

3) TF-IDF and Word2Vec with Neural Network:

- Incorporated TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec embeddings as input features.
- Preprocess text data by tokenizing, vectorizing, and converting it into numerical representations.
- Experimented with different embedding dimensions and vocabulary sizes.

- Combined text embeddings with additional features like category and rating.
- Trained neural networks on the combined feature set and evaluate their performance.

4) LLM (Large Language Model) BERT with Accuracies:

- Utilized pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) for text classification tasks.
- Fine-tuned BERT on the specific task of fake review classification.
- Tokenized input text and encode it using the BERT tokenizer.
- Concatenated BERT embeddings with additional features like category and rating.
- Trained the model on the combined feature set and evaluate its performance using accuracy metrics.
- Compared the accuracy achieved by the BERT model with other approaches to assess its effectiveness in detecting fake reviews.

The BERT model obtained a test accuracy of 94.83% for the Amazon dataset, while the neural network model had the maximum accuracy of 92.98%.

VI. Conclusion

The research focused on hyperparameter tweaking and algorithm assessment to optimize false review detection models using Yelp and Amazon datasets. Amazon's accuracy was 91.74% with a learning rate of 0.001 across 5 epochs, indicating a preference for simpler models. And also, LinearSVC and Ridge Classifier did well on Yelp, with an F1 score of 0.945 and an accuracy of 89.7%, demonstrating the potency of simpler models.

On Yelp, more tests were conducted. A neural network with a learning rate of 0.001 and 15 epochs was trained, resulting in an accuracy of 83.04%. Conventional models were shown to outperform downsampled and upsampled datasets. Although word count categories, text, rating, emotion, and other features were used, BERT did not considerably improve performance. The accuracy of the deep learning model was 83.04%, whereas the original data, which had been upsampled, was 94.83%.

Traditional neural networks, hyperparameter tweaking, Word2Vec, TF-IDF, and BERT models were investigated for Amazon, and the results showed an accuracy of 94.83% on the upsampled original data. In general, the results highlight how crucial it is to use customized strategies and comprehend dataset quirks to effectively detect false reviews.

External Links

- Yelp traditional methodologies
- Amazon traditional methodologies
- Yelp approached methodologies
- Amazon approached methodologies
- Balancing Yelp dataset
- Project Pitch
- Check-point 1
- Check-point 2

References

[1] A. Sihombing and A. C. M. Fong, "Fake Review Detection on Yelp Dataset Using Classification Techniques in Machine Learning," 2019 International Conference on contemporary Computing and Informatics (IC3I), Singapore, 2019, pp. 64-68, doi: 10.1109/IC3I46837.2019.9055644.

[2] Choi, W., Nam, K., Park, M., Yang, S., Hwang, S., & Oh, H. (2023). Fake review identification and utility evaluation model using machine learning. Frontiers in Artificial Intelligence, 5, 1064371. https://doi.org/10.3389/frai.2022.1064371

[3] He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., & Tosyali, A. (2022). Detecting fake-review buyers using network structure: Direct evidence from Amazon. Proceedings of the National Academy of Sciences, 119(47), e2211932119. https://doi.org/10.1073/pnas.2211932119

[4] N. C. Dang, M. N. Moreno-Garc´ıa, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics, vol. 9, no. 3, p. 483, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805,2018.

[6] Yelp Dataset Link: https://github.com/yashpandey474/Identification-of-fake-reviews/blob/main/Datasets/Yelp%20Dataset%20Reduced.csv

[7] Amaon Dataset Link: https://osf.io/tyue9/

[8] ASalminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771. https://doi.org/10.1016/j.jretconser.2021.102771