Detecting Fake Reviews across E-commerce Websites A Comparative Machine Learning Study

Fateen Ahmed
Illinois Institute of Technology
A20545670
fahmed22@hawk.iit.edu

Syed Wali Uddin Quadri
Illinois Institute of Technology
A20554645
suddinquadri@hawk.iit.edu

Mohammed Riyaz Ahmed
Illinois Institute of Technology
A20547233
rmohammed4@hawk.iit.edu

Abstract

The rise of AI-generated fake reviews presents a significant challenge to e-commerce platforms, necessitating robust detection methods. This study presents a comprehensive machine learning-based approach for detecting such reviews, which makes use of a balanced dataset of real Amazon reviews and fake ones created by *GPT-3* and human authors. It compares the effectiveness of various machine learning algorithms, such as *Logistic Regression*, *Naive Bayes*, *SVMs*, *Gradient Boosting*, *Decision Trees*, *Random Forest*, and a *deep learning model*, in distinguishing fake reviews based on linguistic patterns. The results show that feature engineering and dataset balancing are effective in achieving high accuracy across models. The study emphasizes AI's dual role in creating and detecting fake content, emphasizing the importance of advanced detection methods in future research.

1 Introduction

As businesses continue to transition towards digital platforms, e-commerce has witnessed a remarkable growth in user engagement and transaction volumes. With numerous products accessible online, consumers often depend on reviews and ratings as pivotal indicators of a product's quality and utility. Consequently, online reviews have assumed a critical role in the e-commerce ecosystem. However, this reliance has also opened avenues for manipulation through various deceptive practices, posing a significant challenge for platforms like Amazon. A recent and concerning development in this challenge is the emergence of advanced Large Language Models (LLMs) such as ChatGPT. These cutting-edge technologies, adept at generating human-like text, are being leveraged to fabricate reviews, thereby eroding the authenticity of feedback on e-commerce websites. Alarmingly, it has been reported that the current digital marketplaces are riddled with fake reviews, estimated to influence up to a third of consumer purchasing decisions [6]. In light of this, it becomes imperative to devise methods for identifying these AI-generated fake reviews.

In response to this challenge, numerous e-commerce companies, such as Amazon, have turned to machine learning models to discern and counter such fake content[8]. This paper seeks to extend these efforts by examining a range of accessible machine learning models to detect AI-generated fake reviews and evaluating their effectiveness. Our study includes an exploration of popular models, assessing their proficiency in identifying both AI and manually created fake reviews.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1.1 Related Work

In light of the recent advancements in large language models (LLMs), the academic community has been actively studying impact of such models in both positive and negative aspects. While LLMs like GPT-3 offer significant benefits, there is growing concern about their potential role in spreading misinformation. This concern is reflected in the breadth of research dedicated to understanding and mitigating the negative effects of LLMs. Among these studies, two notable works stand out in their relevance to our research focus.

In the first paper, "Combating Misinformation in the Age of LLMs: Opportunities and Challenges," [3] explore the complex role that LLMs play in the creation and identification of fake reviews. This study draws attention to a crucial problem in LLMs called "hallucinations," in which the models create data that is not supported by facts. Interestingly, the paper also suggests that these same LLMs can be effectively used to identify and combat misinformation of this kind. The authors advocate for a balanced approach that leverages the strengths of LLMs while addressing their inherent limitations in the context of misinformation.

The subsequent research, titled "Combat AI With AI: Counteract Machine-Generated Fake Restaurant Reviews on Social Media," offers a comprehensive review of machine learning techniques for detecting fake reviews[5] . In order to determine the legitimacy of online reviews, the paper describes a variety of detection methods, ranging from simple to complex. It describes a number of markers—like the unique writing styles, author profiles, and the chronological nature of reviews—that can be used to distinguish between authentic and fake reviews. Collectively, these studies make a substantial contribution to our comprehension of the challenges associated with detecting fake reviews and highlight the continuous conflict between the advancement of technology and the demand for trustworthy information.

1.2 Problem Statement

The scope of our project includes online businesses that depend on review and rating platforms. In addition to providing information for recommendation models and helping to analyze the consumer's behavior for customized deals and promotions, reviews are essential markers of customer satisfaction with products. It is crucial to verify the legitimacy of these reviews since fake ones have the potential to drastically skew opinions on products and services and affect poor business choices.

As LLM models generate text with human-like features, the traditional detection methods based on temporal analysis and verified user reviews become less effective. Thereby, the new language-based problem demands an integrated approach combining natural language processing (NLP) and machine learning (ML) techniques.

In order to determine the effectiveness of simple classification-based machine learning models in identifying fake reviews, our research uses word frequency methods and NLP-based preprocessing to identify linguistic patterns. We evaluate these models' ability to distinguish real reviews from artificial intelligence-generated fake reviews, concentrating on finding the distinct textual patterns that indicate fake content.

2 Methods

2.1 Dataset Description

2.1.1 Original Dataset: Datafiniti Amazon Consumer Reviews

Our study utilizes the "Datafiniti Amazon Consumer Reviews of Amazon Products" dataset, comprising 5,000 rows and spanning 24 distinct features. This dataset is a rich repository of consumer reviews, offering insights into customer sentiments and opinions about various Amazon products. The key features include:

- Product identification and categorization details (e.g., id, asins, name, brand, categories, primaryCategories).
- Review-specific information (e.g., reviews.rating, reviews.text, reviews.title, reviews.date, reviews.username).
- Additional metadata (e.g., dateAdded, dateUpdated, imageURLs, manufacturer).

This dataset serves as a foundational element for our analysis, providing genuine consumer reviews that are essential for training and evaluating our machine learning models.

2.1.2 Generated Dataset: Balanced Fake Reviews

We also generated a balanced fake reviews dataset using OpenAI's GPT-3 (Da Vinci-3) model. Given API constraints, we were able to generate only 2,000 fake reviews. Consequently, we sampled 2,000 records from the original Datafiniti dataset to maintain balance. The generation process involved:

-> Utilizing the original dataset to maintain consistency in product categories.
-> Generating reviews for each product category across different star ratings (1 to 5 stars), ensuring a diverse and realistic range of feedback.
-> Labeling these reviews as 'fake' for differentiation and analysis purposes. This method allowed us to create a comprehensive set of AI-generated reviews, mimicking various sentiment levels and providing a robust dataset for testing the effectiveness of our machine learning models.

2.1.3 Human-Generated Fake Reviews

In addition to AI-generated reviews, we also utilised human-generated fake reviews. Our team, comprising three members, was tasked with writing fake reviews for known product categories in the original dataset. After creating several hundred reviews, we identified common patterns in our writings. Utilizing these patterns, we replicated the process to produce a sizable number of fake reviews, ensuring a balanced representation of human-generated deceptive feedback.

This approach to generating human-authored fake reviews adds an additional layer to our dataset, offering a more nuanced understanding of how different types of fake reviews can be detected and differentiated from genuine reviews.

Feature Name	Description
id	Unique identifier for each record
dateAdded	Date when the record was added
dateUpdated	Date when the record was last updated
name	Product name
asins	Amazon Standard Identification Numbers
brand	Brand name
categories	Product categories
primaryCategories	Primary category of the product
imageURLs	URLs of product images
keys	Database keys
manufacturer	Product manufacturer
manufacturerNumber	Manufacturer's number
reviews.date	Date of the review
reviews.dateAdded	Date when the review was added
reviews.dateSeen	Date when the review was seen
reviews.doRecommend	Whether the reviewer recommends the product
reviews.id	Unique identifier for each review
reviews.numHelpful	Number of users who found the review helpful
reviews.rating	Rating given by the reviewer
reviews.sourceURLs	Source URLs of the review
reviews.text	Text content of the review
reviews.title	Title of the review
reviews.username	Username of the reviewer
sourceURLs	General source URLs

Table 1: Features of the Original Dataset

Feature Name	Description
primaryCategories	Primary category of the product
reviews.rating	Rating given by the generated review
reviews.text	Text content of the generated review
label	Label indicating the review is fake
reviews.title	Generated title for the product review

Table 2: Features of the Generated Dataset

2.2 Preprocessing

The preprocessing of our dataset began with addressing reviews.date field which was formatted into a standard date and time format for consistency. Also, another crucial step was the elimination of unnecessary features. We removed features that were not useful to our model's objectives, including id, dateAdded, dateUpdated, and such. These features, comprising mostly of various identifiers and URLs, were irrelevant for our modeling purposes. Next, we transformed the remaining relevant features. The reviews.doRecommend feature, consisting of True and False values, was interpreted as an indicator of real (True) and fake (False) reviews. We mapped these to numerical values of 0 (real) and 1 (fake) for simplicity. The PrimaryCategories feature, encompassing four categories—Electronics, Hardware, Media, and Office Supplies—was numerically encoded for model compatibility.

In the feature creation phase, we focused on the reviews.text attribute, deriving several key features:
Word Count: We categorized review lengths as 'short' (less than 10 words), 'moderate' (10 to 50 words), and 'long' (over 50 words).
Sentiment Analysis: Using the TextBlob library, we classified the sentiment of each review as either positive or negative.
Frequency of Words: This metric indicated the repetition of specific words across reviews, classified into low, moderate, and high frequency categories.

The generation of fake reviews was integral to our project, aiming to detect AI-generated and human-generated fake reviews. After producing these fake reviews, we merged them with the real reviews, labeling them as 0 (real) and 1 (fake). We then balanced our dataset to prevent bias towards real or fake reviews. The dataset was split into training, testing, and cross-validation sets, allocating 80 percent for training and 20 for testing.

The final step in preprocessing involved the application of two critical text vectorization techniques:
TF-IDF Vectorization: This method weighs the frequency of words in each document against their frequency across the entire corpus, helping to highlight words that are unique to specific documents.
Bag of Words (BoW): This simpler approach transforms text into a numerical representation based on word frequency, disregarding the order or context of words. Both these techniques were employed to convert the textual content of reviews into a numerical format suitable for machine learning algorithms. Our dataset, post-preprocessing, comprised 4,765 real reviews and 4,343 fake reviews, ensuring a balanced distribution for effective model training and testing

2.3 Model Selection and Implementation

This section covers the information related to model selection and its implementation. We chose simple but effective models. These models are not only easy to understand, but they also work really well. Since our project consists of two objectives: detecting fake reviews generated by AI and also detecting human-generated fake reviews, both will be discussed.

2.3.1 Classical Machine Learning Methods

To tackle the problem of identifying fake reviews, we utilized multiple traditional machine learning models, each selected based on their suitability to the field of study.

Logistic Regression:
Because of its effectiveness in differentiating between two classes, *logistic regression* is a mainstay in binary classification tasks and is therefore ideal for classifying reviews as authentic or fake. This model is very good at giving probability-based results, revealing information about how likely it is that a review is genuine. Its computational efficiency was a key consideration in its inclusion, given the large volume of online reviews.

Implementation:

- *Hyperparameters:*
  - TF-IDF Vectorization: Max Features = 50, Exclude Stop Words = English
  - Logistic Regression: Initialized with default hyperparameters.
- *Process:*
  1. Data Manipulation and Analysis ($\mathcal{D}_{manip}$):
     - Use of pandas library for data manipulation ($pandas_{lib}$).
     - CSV data split: Training Set = 80%, Testing Set = 20%.
  2. Feature Extraction ($\mathcal{F}_{ext}$):
     - Application of TF-IDF vectorization ($TF-IDF_{vec}$) to text reviews.
     - Transformation of dataset into a matrix of TF-IDF features ($Matrix_{TF-IDF}$).
  3. Defining Target Variable ($\mathcal{T}_{var}$):
     - 'label' column designated as target variable for training (y_{train}) and test datasets (y_{test}).
  4. Model Training and Validation ($\mathcal{M}_{train}$):
     - Logistic Regression model trained with TF-IDF features (X_{train}) and corresponding labels (y_{train}).
     - Cross-validation (CV_{val}) employed to assess model performance.

Naive Bayes:
As the task involves text data, Naive Bayes is an easy to implement and efficient in handling such data. Moreover, its probabilistic approach and strong text classification performance make it an good choice for identifying the linguistic patterns typical of fake reviews. The model also works well on fewer data.

Implementation:

Hyperparameters:

- *CountVectorizer:*
  - $max\_features = 50$
  - $stop\_words = English$
- *Multinomial Naive Bayes:*
  - $\alpha = 1.0$ (Additive smoothing parameter)
  - $fit\_prior = True$ (Learn class prior probabilities)

Process:

- Read training and test datasets using pandas.
- Extract 'label' as target variable.
- Convert text reviews into BOW representations using CountVectorizer.
- Limit features to 50, excluding English stop words.
- Initialize and train Multinomial Naive Bayes model.

Support Vector Machines:
Support Vector Machines is another classical approach that is utilized. It works on the basis maximizing the margin between classes to ensure a robust and generalizable decision boundary. We test this model too to check if it works well in detecting the fake reviews.

Implementation:

- *Hyperparameter:*
  - The SVM model employs a linear kernel (kernel='linear').
- *Implementation Process:*
  1. Model Initialization:
     - Initialized an SVM classifier (SVC) with a linear kernel and a random seed for reproducibility.
  2. Model Training:
     - Fitted the SVM model to the training data (Xtrain, ytrain).

Gradient Boosting Classifier:
Gradient boosting based techniques offers a choice over various loss functions and the provides adaptability. Moreover, it uses several hyper parameters that can be tuned to improve performance. It is also resistant to overfitting as it incorporates techniques such as early stopping. We also utilized gradient boosting in our research due to such properties.

Hyperparameters:

- Number of Estimators ($n\_estimators$): 100
- Learning Rate ($learning\_rate$): Set to 0.1
- Maximum Depth (max_depth): 3,
- Minimum Samples Split ($min\_samples\_split$): 2
- Minimum Samples Leaf ($min\_samples\_leaf$): 1
- Subsample ($subsample$): 1.0

Decision Tree Classifier: We also selected to implement the decision tree classifier as its a foundation to ensemble methods which will be discussed subsequently. The decision tree classifier is robust to outliers and is reliable and easy to implement model which we wanted to compare with other classical models.

2.3.2 Ensemble Method

As the paper's aim is to compare various machine learning models, it is important to include various types of machine learning, one such method is the ensemble approach. The random forest model is implemented in order to classify the given task. It is known for its accuracy and resilience, particularly in scenarios with imbalanced datasets, the Random Forest Classifier was employed to enhance the robustness of our detection mechanism. Its built-in out-of-bag error estimation provides an internal validation mechanism, a valuable asset for ensuring the reliability of our predictions. This technique also reduces the risk of overfitting, by averaging out the biases across individual trees, ensuring a more generalizable model.

Random Forest Implementation:

- *Hyperparameters:*
  - Number of estimators: 5 (n_estimators=5)
  - Random state for reproducibility: 42 (random_state=42)
- *Process:*
  1. Model Initialization and Training:
     - Initialized a Random Forest classifier with 5 estimators and a random state of 42.
     - Fitted the model to the TF-IDF transformed training data.
     - Made predictions on the test set.
     - Evaluated the model's performance using classification report metrics like precision, recall, and F1-score.

2.3.3 Deep Learning Method

Neural networks are another powerful subset of machine learning models that learn from high-dimensional data and find complex relationships between the various parameters. It can also be utilized to detect fake reviews, and in this study, we implemented the feed-forward neural network, which utilizes multiple layers and nonlinear calculations to capture the relationships in the data. It is also well suited to handle high-dimensional data, which usually occurs when we utilize vectorizers to translate text with a larger vocabulary set.The deep learning approach in our study involved the implementation of a feedforward neural network for text classification.

Neural Network Architecture:
A feedforward neural network, named *TextClassifier*, was defined inheriting from the nn.Module class in PyTorch. The network architecture includes:

- Three fully connected (linear) layers with ReLU activation functions.
- An input layer size corresponding to the TF-IDF feature vector size (max_features=50).
- A single neuron in the output layer with a sigmoid activation function for binary classification.

Model Training:

- Initialized the model along with the binary cross-entropy loss function (BCELoss) and the Adam optimizer.
- Conducted training over 5 epochs using mini-batch gradient descent.
- Computed loss, performed backpropagation, and updated the model parameters during training.

Hyperparameters:

- Learning rate: 0.001
- Number of epochs: 5
- Batch size: 64
- Model architecture: 3 fully connected layers with neuron counts of 512, 128, and 1 respectively, with ReLU activations.

3 Results

Our models were evaluated based on accuracy, recall, F1-score, and precision metrics. The graphical representation of these metrics provides a clear comparison across different models used in our study.

We find that accuracy, precision, recall, and F1-score are all consistently high when we examine the data from the various models. These findings point to the effectiveness of our preprocessing procedures as well as the distinct nature of the features we took out of the review text. The distinct linguistic patterns in our dataset that distinguish real reviews from fake ones may be the reason behind the effectiveness of the Logistic Regression, Naive Bayes, Random Forest, and neural network models. Preprocessing techniques like TF-IDF vectorization and the Bag-of-Words model probably enhanced the key characteristics of fake reviews, making it easier for the models to identify them.

Further, the ***graphs*** are presented -

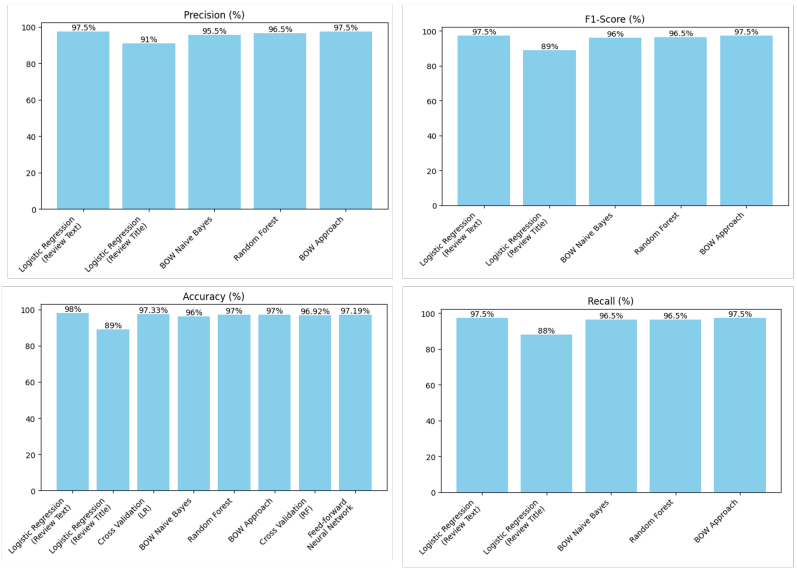

Figure 1: Evaluation metric comparison across different machine learning algorithms.

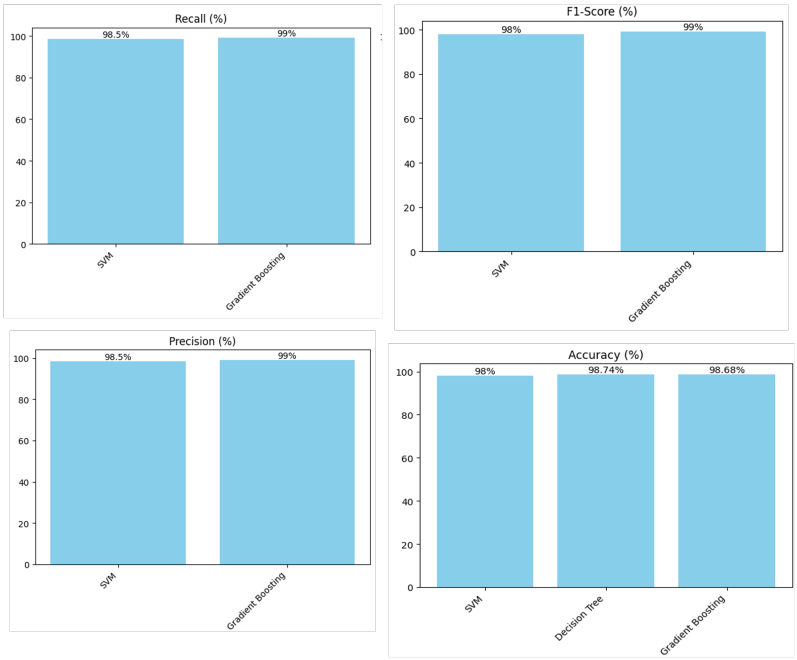

Figure 2: Evaluation metric comparison across different machine learning algorithms.

The balanced nature of the dataset following resampling, which made sure the models weren't biased toward a particular class, may also be partially responsible for the high performance. Furthermore, the models' focus upon linguistically relevant features for fake review detection appears to have improved as a result of the decision to concentrate on textual content and related metadata, rather than outside variables.

However, the high metrics may be a sign of overfitting, particularly if the dataset isn't as complex as real-world data. Moreover, the strong performance may not adapt well to more complex fake reviews that use advanced techniques to more closely resemble real reviews. To make sure that these results hold true in more general applications, future work could explore deep learning models that capture more subtle linguistic patterns and cross-validate with external datasets.

4 Conclusion and Future Work

This study has systematically evaluated the capability of various machine learning algorithms to distinguish between authentic and AI-generated fake reviews. Our results show that both the advanced and classical models are highly accurate, which is due to the effective feature engineering and the balanced the dataset utilsed in the project.

Though encouraging, we acknowledge that these findings might not fully capture the complex nature of the problem at hand. The possibility of overfitting to our particular dataset suggests that, in order to guarantee the generalizability of the models, more extensive validation against more complicated and diverse datasets is required.

The future work includes extending the scope of linguistic analysis, utilizing unsupervised models for anomaly detection, and integrating analytics related to user behavior. Such multimodal approaches might reveal more complex and subtle strategies that our current models might miss when creating fake reviews.

As the prevalence of artificial intelligence (AI)-generated fake reviews increases, this study offers relevant information about how well basic machine learning algorithms detect these reviews. Sophisticated AI models are having an increasing impact on many domains, such as artistic content and consumer reviews, even though ethical considerations in AI development are still developing. As noted in the related work section, this study emphasizes the dual role of artificial intelligence (AI)—not only in producing misleading content but also in its capacity to identify and prevent such actions.

References

[1] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] Brightmart, "Text Classification Models." [Online]. Available: `https://github.com/brightmart/text_classification`

[3] Canyu Chen and Kai Shu, "Combating Misinformation in the Age of LLMs: Opportunities and Challenges," *arXiv preprint arXiv:2311.05656*, 2023. [Online]. Available: `https://arxiv.org/pdf/2311.05656.pdf`

[4] "Datafiniti Amazon Consumer Reviews of Amazon Products," Datafiniti. [Online]. Available: `https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products`

[5] Alessandro Gambetti and Qiwei Han, "Combat AI With AI: Counteract Machine-Generated Fake Restaurant Reviews on Social Media," *arXiv preprint arXiv:2302.07731*, 2023. [Online]. Available: `https://arxiv.org/pdf/2302.07731.pdf`

[6] "Fake reviews: how AI is creating a new battleground for hotels, restaurants and products," *The Guardian*, July 15, 2023. [Online]. Available: `https://www.theguardian.com/money/2023/jul/15/fake-reviews-ai-artificial-intelligence-hotels-restaurants-products`

[7] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[8] "How AI spots fake reviews on Amazon," Amazon, 2023. [Online]. Available: `https://www.aboutamazon.com/news/policy-news-views/how-ai-spots-fake-reviews-amazon`

The complete source code for our project, titled "Research Project on AI-Generated Fake Reviews," can be accessed on GitHub at the following URL:
https://github.com/alfa2k/Research-Project-On-AI-Generated-Fake-Reviews.git