AI 생성 텍스트 탐지: 가짜 리뷰 식별을 위한 전통 머신러닝과 BERT의 비교 분석

Detecting AI-Generated Text: A Comparative Analysis of Traditional Machine Learning and BERT for Fake Review Identification

2022203510 유니버스

Abstract: AI-generated text can be used to create fake online reviews, which misleads consumers and erodes trust in e-commerce. This paper compares different methods for detecting these fake reviews. Traditional machine learning models such as Logistic Regression and Support Vector Machine (SVM) were tested against a modern deep learning model, BERT. The models were evaluated on a dataset of real Amazon reviews and fake reviews generated by GPT-2. The results show that BERT is significantly more effective, achieving 97.4% accuracy, while the traditional models reached about 93.6%. This work confirms that deep learning models that understand context are better suited for identifying the subtle patterns in AI-generated text

Keywords: BERT, fake review detection, GPT-2, binary classification, transformers, text classification, NLP

1. Problem Definition

AI models like GPT-2 can write product reviews that are highly realistic. When used maliciously, these fake reviews can mislead customers and damage the credibility of online platforms. The central problem this research addresses is the **accurate detection of AI-generated product reviews** to distinguish them from those written by humans. If this problem is not solved, consumers may make poor purchasing decisions based on fraudulent content, and online platforms may lose user trust.

Previous detection methods often rely on traditional machine learning models with engineered text features, such as word frequency counts (TF-IDF). These approaches struggle to capture the subtle linguistic patterns of AI-generated text because they do not fully account for the context of words and sentences. This research fills that gap by directly comparing these traditional methods to a state-of-the-art deep learning model, BERT, which is specifically designed to understand language context. The goal is to determine which approach is more effective for this detection task.

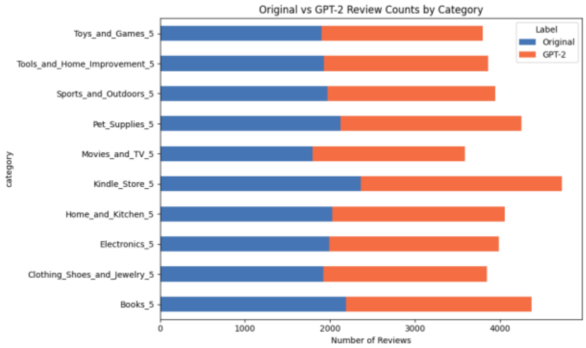
2. Dataset Description

The models in this study were trained and evaluated on the "Fake Reviews Dataset" sourced from Kaggle, which is structured for a binary classification task. This dataset combines genuine human-written reviews with computer-generated fakes.

Data Source: The dataset contains genuine product reviews from Amazon and synthetic reviews generated by the GPT-2 model.

Labels: Each review is labeled as either Original (human-written) or GPT-2 (AI-generated). Originally in the dataset, it was labelled as OR = Original reviews (presumably human created and authentic) and CG = Computer-generated fake reviews.

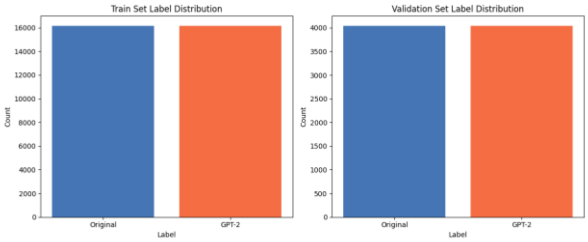
[Fig1]: Dataset Composition by Product Category



Preprocessing: A minimal preprocessing was performed to preserve the text's natural style. All text was converted to lowercase. For the traditional models, the character length of each review was also calculated and this value was normalized using a MinMaxScaler.

Data Splitting: The dataset was divided into an 80% training set and a 20% validation set. A stratified split was used to ensure that the ratio of real to fake reviews was identical in both sets, which prevents class imbalance from affecting the results.

[Fig2]: Label Distribution in Training and Validation Sets.



3. Proposed Method

This study compares two distinct approaches to the detection problem: a traditional, feature-based method and a deep learning method.

3.1. Approach 1: Traditional Machine Learning Models

The baseline approach uses two standard machine learning classifiers: Logistic Regression and a Support Vector Machine (SVM). These models depend on features that was manually engineer from the review text.

Feature Engineering: The text was converted into numerical features using two components:

a. TF-IDF Vectorization: Term Frequency-Inverse Document Frequency (TF-IDF) was used to represent the text. This method was configured to analyze both single words (unigrams) and two-word phrases (bigrams), with the vocabulary limited to the 5,000 most frequent features.

b. Feature Combination: The resulting TF-IDF matrix was combined with the normalized review length feature.

Training: Both models were implemented using the scikit-learn library and trained with the max_iter parameter set to 1000 to ensure convergence.

3.2. Approach 2: Fine-Tuned BERT Model

The second approach uses a pre-trained Transformer model, BERT (Bidirectional Encoder Representations from Transformers). This method does not require manual feature engineering, as the model learns to understand text context directly from the raw data.

Architecture: TFBertForSequenceClassification model was used from the HuggingFace Transformers library, which is pre-built for classification tasks.

Tokenization: Text was processed with the BertTokenizer. All reviews were either padded or truncated to a uniform length of 64 tokens.

Training: The model was fine-tuned on a Google Colab GPU using TensorFlow. Key training parameters were:

Optimizer: Adam
Loss Function: SparseCategoricalCrossentropy
Batch Size: 16
Epochs: 2
Learning Rate: 5e-5

4. Experimental Results and Evaluation

The performance of all models was evaluated on the held-out validation set.

4.1. Evaluation Metrics

As this is a binary classification task, a standard set of metrics was used to provide a comprehensive assessment of performance:

Accuracy: The overall percentage of correct predictions.

Precision: Of all reviews flagged as AI-generated, the percentage that were actually AI-generated.

Recall: The percentage of all AI-generated reviews that were correctly identified.

F1-Score: The harmonic mean of Precision and Recall, offering a balanced measure of a model's performance.

4.2. Overall performance

The experimental results, summarized in Table 1, show a clear performance advantage for the fine-tuned BERT model.

[Table1]: Performance Comparison of Models. The best result for each metric is in bold.

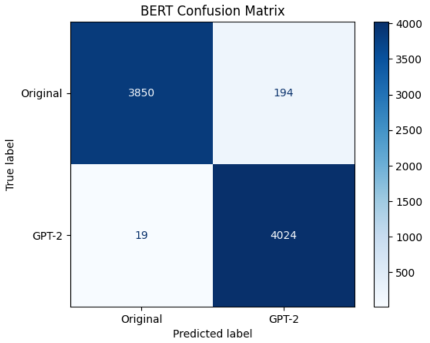
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	93.4%	93.5%	93.3%	93.4%
SVM	93.6%	93.8%	93.5%	93.6%
BERT	**97.4%**	**97.5%**	**97.3%**	**97.4%**

BERT achieved an accuracy of 97.4%, a significant improvement over the SVM (93.6%) and Logistic Regression (93.4%) models. A review of the confusion matrices also confirmed that BERT made fewer errors in both false positives and false negatives.

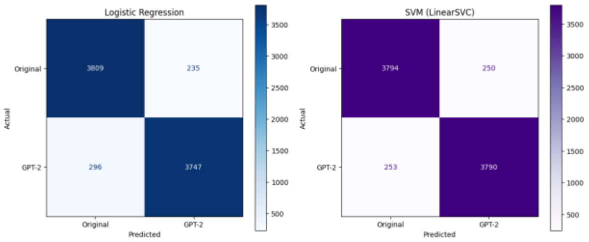
4.3 Error Analysis and Model Comparison

An analysis of the confusion matrices (Figure 2 and Figure 3) confirms BERT's superior performance. The BERT model made significantly fewer errors, especially in misclassifying fake reviews as genuine (only 19 false negatives).

[Fig3]: Confusion Matrix for Fine-Tuned BERT Model

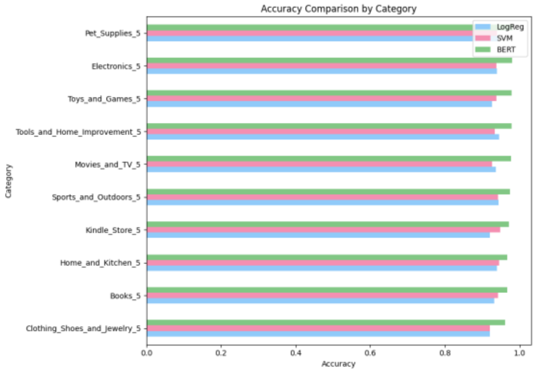


[Fig4]: Confusion Matrices for Baseline Models (Logistic Regression and SVM)



Furthermore, **Figure 4** shows that BERT's high accuracy is consistent across all 10 product categories, unlike the baseline models whose performance varies more significantly between categories.

[Fig4]: Accuracy Comparison by Product Category



4.4 Feature Importance in Baseline Models

To better understand how the traditional models distinguish between human and AI-generated text, a feature importance analysis was conducted. The coefficients assigned to each TF-IDF feature by the Logistic Regression and SVM models reveal which words or phrases are the most predictive. A high positive coefficient indicates a strong predictor for a GPT-2 review, while a large negative coefficient is a strong predictor for an Original, human-written review.

Table 2 and **Table 3** list the top 10 most influential features for the Logistic Regression and SVM models, respectively.

[Table 2]: Top TF-IDF Features for Logistic Regression

LogReg GPT-2	**Value**	**LogReg Original**	**Value**
the only:	6.9501	at:	-5.6429
will keep:	6.2565	even:	-4.9757
and it:	6.0950	but:	-4.7557
and the:	5.0201	in:	-4.7385
but it:	4.3208	though:	-4.5239
that it:	4.1943	on:	-4.5070
has the:	3.9628	from:	-4.3376
had to:	3.9399	no:	-4.1608
this for:	3.6344	all:	-4.1214
the story:	3.6051	to:	-4.0385

[Table 3]: Top TF-IDF Features for SVM

SVM - GPT-2	**Value**	**SVM - Original**	**Value**
will keep:	4.2560	at:	-3.1001
the only:	4.2372	though:	-2.9549
and it:	3.4733	even:	-2.7046
but it:	2.9576	but:	-2.5561
have one:	2.8849	to:	-2.5580
also love:	2.7048	my:	-2.4501
has the:	2.6587	and:	-2.3955
it been:	2.6265	no:	-2.3222
had to:	2.6173	because:	-2.1842
is little:	2.6168	sometimes:	-2.1102

The analysis reveals several interesting patterns. Both models identified similar phrases as strong indicators of AI-generated text, such as the bigrams "the only:", "will keep:", and "and it:". These common conjunctive phrases and declarative statements appear to be stylistic artifacts of the GPT-2 model.

Conversely, words that are highly predictive of original human reviews include common prepositions and conjunctions like "at:", "but:", and "no:", as well as personal pronouns like "my:". The strong consistency in top features across two different models suggests that these are reliable signals for TF-IDF-based classifiers on this task. This contrasts with the BERT model, which relies on deeper contextual understanding rather than specific word frequencies.

4.5. Evaluation and Discussion

The results strongly indicate that deep contextual models are better suited for detecting AI-generated text. BERT's superior performance is due to its ability to understand the semantic context of language, which allows it to identify subtle generative patterns that are missed by feature-based methods like TF-IDF.

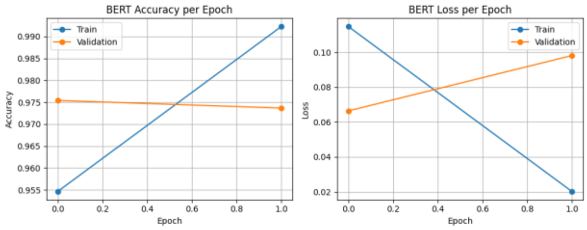
This study has several limitations. First, the dataset was limited to text generated by GPT-2, and the findings may not generalize to newer models like GPT-3 or GPT-4. Second, the analysis was restricted to the review text and did not include other potentially useful signals, such as reviewer metadata. Finally, training BERT is more computationally expensive than training traditional models and requires GPU resources.

In conclusion, this work demonstrates that a fine-tuned BERT model is a highly effective and precise tool for identifying AI-generated reviews. Future work should focus on testing these methods against a broader range of generative models and exploring the use of multimodal data (e.g., text plus user metadata) to further improve detection capabilities.

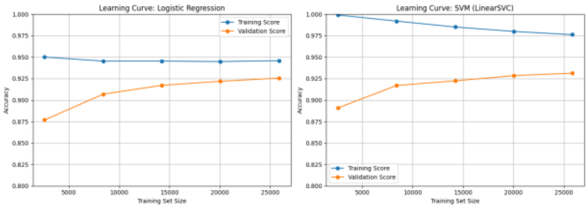
Appendix

A1: Model Training Dynamics

[FigA1]: Accuracy and Loss Curves for BERT Fine-Tuning.



[FigA2]: Learning Curves for Baseline Models (Logistic Regression and SVM).



A2: Qualitative Error Analysis

[TableA1]: Examples of Misclassified Reviews by the BERT Model. This table shows samples where the model's prediction (Pred) did not match the true Label (0 = Original, 1 = GPT-2).

Text (truncated)	**Label**	**Pred**
This socks are simply amazing. So soft and br...	0	1
This is exactly as advertised. Does. To get mo...	0	1
I very well pleased with item and it works ver...	0	1
I ordered “Grover” for my niece for Christmas...	0	1
GREAT READ SOMETHING DIFFERENT SHE WAS MYTHOLO...	1	0

References

[1] J. Salminen, C. Kandpal, A. M. Kamel, S. Jung, and B. J. Jansen, “Creating and detecting fake reviews of online products,” *Journal of Retailing and Consumer Services*, [Online], vol. 64, art. 102771, 2022, DOI: 10.1016/j.jretconser.2021.102771. Available: https://doi.org/10.1016/j.jretconser.2021.102771.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv*, [Online], vol. abs/1810.04805, 2018. Available: http://arxiv.org/abs/1810.04805.

[3] S. De and A. Vats, “Unmask It! AI-Generated Product Review Detection in Dravidian Languages,” *arXiv*, [Online], 2025. Available: https://arxiv.org/abs/2503.09289.

[4] D. Pithadia, S. Pithadia, B. Shah, and M. Bhavsar, “A Comparative Analysis of Machine Learning Models for Fake Review Detection,” *Informatica*, [Online], vol. 47, no. 4, 2023. Available: https://www.informatica.si/index.php/informatica/article/view/7071.

[5] A. Sharma, *HuggingFace*, "bert-base-uncased," [Online], https://huggingface.co/google-bert/bert-base-uncased, Accessed: June 20, 2025.

[6] aayush210789, “Deception-Detection-on-Amazon-reviews-dataset/SVM_model.ipynb,” *GitHub*, [Online], https://github.com/aayush210789/Deception-Detection-on-Amazon-reviews-dataset/blob/master/SVM_model.ipynb, Accessed: June 20, 2025.