engineering proceedings

MDPI

Proceeding Paper

Classification of Artificial Intelligence-Generated Product Reviews on Amazon †

Jia-Luen Yang

Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA; jialuenyang@gmail.com

† Presented at the 2024 IEEE 6th Eurasia Conference on IoT, Communication and Engineering, Yunlin, Taiwan, 15–17 November 2024.

Abstract: Amazon has been flooded with artificial intelligence (AI)-generated product reviews that offer minimal value to customers. These AI reviews merely echo the given product descriptions without providing any authentic information on how buyers feel when using the products. Therefore, an AI review-identifying method was developed to enhance the quality of the review-reading experience in this study. A dataset of 6217 Amazon reviews was compiled including 1116 identified as AI-generated ones. They were classified with a 99.25% F1 score on the test data using the term frequency–inverse document frequency (TF–IDF) and support vector classifier (SVC). The developed method enables the detection of AI-generated reviews on the internet, fostering an authentic and reliable platform.

Keywords: AI content detection; NLP; TF–IDF; SVC; machine learning

1. Introduction

Since the end of 2022, people have increasingly used large language models (LLMs) such as OpenAI's ChatGPT for their writing. Many writers have used LLMs to compose emails or develop outlines for academic papers. While LLMs are beneficial for writing applications, significant concerns also have been raised, particularly regarding product reviews. Amazon has seen a surge in artificial intelligence (AI)-generated reviews following the public release of ChatGPT. CNBC published a news article bringing attention to Amazon product reviews that began with disclaimers such as "As an AI language model" [1]. However, this practice has diminished as LLMs evolve. Modern LLMs generate reviews without explicitly referencing their artificial nature. From the researcher's experience, the first few reviews on any recent product page might have been generated by AI. This poses a problem for potential buyers who rely on reviews to make informed purchasing decisions. AI-generated reviews often lack substantive values and unduly influence buyers by presenting products in an overly favorable opinion. To address this issue, a method was developed in this study for detecting AI-generated product reviews with over 99% accuracy. Various techniques, including token analysis, sentiment analysis, and term frequency–inverse document frequency (TF–IDF) were employed along with classifiers of naïve Bayes, logistic regression, and linear support vector classification to achieve high F1 scores in distinguishing genuine human-written reviews from LLM-written reviews.

Amazon employs AI to verify the authenticity of product reviews submitted by customers [2]. They do not announce what techniques or methods are used to detect fake reviews. However, recent product reviews on Amazon reveal that a significant number

check for updates

Academic Editors: Teen-Hang Meen, Chi-Ting Ho and Cheng-Fu Yang

Published: 25 April 2025

Citation: Yang, J.-L. Classification of Artificial Intelligence-Generated Product Reviews on Amazon. *Eng. Proc.* **2025**, *92*, 17. https://doi.org/10.3390/engproc2025092017

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

of reviews are AI-generated. This raises the question of whether Amazon can classify AI-generated reviews.

Potential buyers often check the reviews of the products they intend to buy. The data gathered in this study showed that the top 100 recent product reviews contained up to five or six AI-generated reviews. Notably, many of these AI-generated reviews were ranked in the top three which were displayed on the products' default pages. These reviews generally lacked authenticity and did not provide useful information for buyers; they reiterated details already included in the product description without offering valuable information on the products. Dugan indicated that humans have the cognitive capability to consistently identify AI-generated sentences with practice [3]. However, the process takes time and effort, which many people are unwilling to invest. Therefore, it is essential to detect and remove AI-generated reviews to improve the product-browsing experience and ensure that potential buyers can focus on genuine, insightful reviews written by humans.

There are paraphrasing tools that can rephrase AI-generated content to evade detection. Though these tools bypass LLM watermarking documents [4], this is not a concern for product reviews. The reviewers who use LLMs can save their time and effort. They mostly copy and paste the generated paragraphs into Amazon. In several cases, reviewers even accidentally copied their entire conversations with ChatGPT. Much research has been conducted regarding AI-generated content detection. Most studies focus on the forms and varieties of texts included in emails, essays, papers, etc. In contrast, product reviews have not been studied extensively. The results of this study enable the development and implementation of finetuned and efficient methods for distinguishing between texts written by AI and humans.

2. Related Work

Many studies have been conducted on detecting AI-generated texts. There are differences between humans and ChatGPT in writing. Humans are more likely to be colloquial [5]. It is easier to detect AI-generated full paragraphs than a single sentence due to ChatGPT's use of repeated keywords. This information was used for the selection of tools for detecting AI-generated texts in this study. Tang, Chuang, and Hu researched black-box and white-box detections [6]. They acknowledged that the TF–IDF technique was useful in detecting texts generated by GPT-2. Nowadays, people use more advanced GPT-3.5 and -4 to write product reviews. As these LLMs have become more sophisticated than before, a more refined method is required to detect AI-generated content in all forms of text. However, in this study, product reviews were explored since the keywords and phrases GPT-3.5 and -4 output exhibit significant overlaps and similarities which does not necessitate a sophisticated method.

Bhattacharjee and Liu used the idea of fighting fire with fire by using ChatGPT to detect AI-generated content [7]. While this approach is promising, it relies on OpenAI's online service. In this study, the method to detect AI-generated texts quickly and efficiently without depending on third-party services was developed. The method is capable of running even on older hardware. GPT-4 misclassified texts generated by ChatGPT as human-written ones 50% of the time [7]. The accuracy of GPT-4 is too low to be used as a classifier.

Prova explored various techniques and methods such as CountVectorizer, TF–IDF, bidirectional encoder representations from transformers (BERT), extreme gradient boosting (XGB), and support vector machine (SVM) [8]. While such techniques were similar to the developed method in this study, the results are different due to the dataset. Prova created a dataset with 3000 data points using ChatGPT but did not disclose the types of prompts

used. In contrast, existing product review texts were collected on Amazon and then labeled as AI-generated or human-written for this study.

Salminen et al. created a fake Amazon product reviews dataset [9]. They trained GPT-2 to output fake reviews by feeding it tens of thousands of human-written Amazon reviews. The product reviews created by GPT-2 were better at mimicking human writing than the AI-generated reviews found on Amazon. As potential buyers do not have the knowledge or tools to train their LLMs, such a dataset does not accurately represent the AI-generated reviews on Amazon.

3. Materials and Methods

The dataset of Amazon product reviews was created by scraping, preprocessing, and manual labeling. Distinct patterns in vocabulary and style between the two types were analyzed. Various NLP techniques and classifiers were employed to identify AI-generated texts based on features including common keywords, sentiment, and linguistic characteristics.

3.1. Data Collection

The dataset contained product reviews collected through data scraping, preprocessing, and manual labeling in this study. Amazon reviews and other metadata containing images or videos were collected. Then, the collected data were preprocessed to remove reviews in other languages than English. Finally, each review was manually read and labeled. The dataset comprised 6217 reviews, of which 1116 were AI-generated.

For preprocessing, the Python package langdetect was used to determine whether a review was written in English. The reviews in other languages were removed from the dataset. Reviews written in multiple languages, such as English and Spanish, were classified as reviews in English to maintain consistency. Each review was labeled as 0 for a human-written review and 1 for an AI-generated review. Labeling was conducted based on the researcher's experience. Several reviews were written by humans with the help of AI. The writer used LLMs to organize and elaborate on his or her thoughts. These AI-assisted reviews contained details regarding product use. There were 145 AI-assisted written reviews in the dataset. For simplicity, these reviews were classified as AI-generated ones. The entire dataset is publicly available on the researcher's GitHub repository.

3.2. Data Analysis

The review texts of each label were analyzed to determine the most common tokens per label. Stop words were removed to reduce computational load. Each token was counted only once even if it was used multiple times in the same review. Human-written reviews contained common positive words such as "great" and "good". They also included emphasis on how easy or convenient it was to "use" the products. Even the most common word "great" appeared in less than 30% of all human-written reviews (Table 1).

Table 1. Top five most common human-written tokens.

Tokens	Frequency	Percentage of Human-Written Reviews
great	1365	26.8%
use	1111	21.8%
like	1094	21.4%
good	1031	20.2%
Just	733	14.4%

Table 2 illustrates the lack of diverse vocabulary in AI-generated reviews. The top three most common tokens were used in 50% of the reviews. AI-generated reviews tended

to present different aspects of the product ranging from its appearance to its condition with frequent use of "design" and "quality". AI-generated reviews were formatted as academic essays with a short conclusion to stress the previous opinions, and "overall" was commonly used.

Table 2. Top five most common AI-generated tokens.

Tokens	Frequency	Percentage of AI-Generated Reviews
design	579	51.9%
making	551	49.4%
quality	524	47.0%
overall	491	44.0%
use	422	37.8%

3.3. Experiment

NLP techniques such as TF–IDF, sentiment analysis, and token analysis were employed to identify AI-generated texts in this study. These methods are efficient on the text. Additionally, classifiers, including multinomial naïve Bayes, complement naïve Bayes, logistic regression, and linear support vector classification were used. AI-generated texts frequently included specific keywords and phrases such as "elevated", "game-changer", and "exceeded all my expectations". Those keywords and phrases were positive remarks for products. TF–IDF enabled assessing how closely a review resembled typical AI-generated ones. If keywords commonly found in AI-generated reviews were included, a greater chance of the review being written by an LLM was assumed. In contrast, human-written reviews tended to include various vocabulary with an emphasis on the product details and specifics.

Sentiment analysis was employed because most AI-generated reviews were overwhelmingly positive and filled with praises. Gillham stated that extreme reviews on Amazon were found 1.3 times more in AI-generated reviews than human-written ones [10]. The reviewers used LLMs to generate positive reviews based on the product descriptions. Therefore, the more positive the reviews were, the more likely they were written by AI. Sentiment analysis was used to calculate a ratio of positive, negative, and neutral sentiments for any given review. The ratio was then used as a feature to train the classifiers.

Token analysis was used to determine features such as token length, spelling errors, and grammatical mistakes, which are valuable indicators of human-written reviews. Most publicly available LLMs were trained to give outputs with perfect spelling and grammar, reducing the likelihood of producing AI-generated texts with spelling errors. The presence of spelling errors, therefore, was an indicator of human-written reviews.

4. Results

TF–IDF, sentiment analysis, and token analysis were trained separately on the training dataset using different classifiers. Then, the classifiers were evaluated based on the test dataset. TF–IDF yielded better results than sentiment analysis and token analysis. SVC showed the highest F1 score for each feature (Table 3).

Table 3. F1 scores based on each feature.

Features	F1 Score
TF–IDF	0.9925
Sentiment analysis	0.8548
Token analysis	0.2365

SVC achieved the highest F1 score of 0.9925 on the test dataset. Naïve Bayes and multinomial naïve Bayes demonstrated the lowest accuracy among all the classifiers. Logistic regression showed the second-best performance (Table 4).

Table 4. TF–IDF F1 scores based on each classifier.

Classifiers	F1 Score
Complement naïve Bayes	0.9673
Multinomial naïve Bayes	0.9352
Logistic regression	0.9732
Support vector classifier	0.9925

TF–IDF with SVC misclassified only one human-written review (Table 5). This result demonstrated that the likelihood of accidentally classifying human-written reviews as AI-generated was extremely low. Misclassified AI-generated reviews were less penalized than other AI-generated reviews. The human-written reviews were retained as much as possible while eliminating the majority of AI-generated ones.

Table 5. TF–IDF SVC confusion matrix.

		Predicted	
		Human	**AI**
True	**Human**	1530	1
	AI	13	322

5. Discussion

The best results were achieved when using only the TF–IDF feature with SVC. TF–IDF effectively measured how closely a review matched the characteristics of AI-generated reviews. The effectiveness of TF–IDF is owing to the widespread use of ChatGPT. Most AI-generated product reviews were written by ChatGPT, which was the most popular public LLM. The outputs of ChatGPT contained similar keywords and phrases with the command "Write a positive review for this product". Such a correlation was evident in the manual review of the dataset which was subsequently confirmed through TF–IDF.

Sentimental and token analyses did not improve the results as they could not classify AI-generated texts better than TF–IDF. While most AI-generated reviews were predominantly positive, several presented negative sentiments. Some AI-generated reviews contained a mixture of positive and negative sentiments resembling human-written reviews. A human reviewer typically discussed the pros and cons of a product, resulting in blended positive and negative sentiments.

The developed method in this study presents the rapid processing of large volumes of data with TF–IDF and SVC. However, the misclassification of AI-generated reviews occurred. Such reviews were written from a different perspective using unique vocabulary, which was different from the TF–IDF keywords. Addressing this issue requires a larger dataset to improve the accuracy of classification. It is also necessary to develop methods to distinguish between human-written and AI-generated sentences within the same product review. When reviewers used LLMs with bullet points outlining the pros and cons of products, such AI-assisted human reviews were classified as AI-generated reviews in this study. Although these reviews were generated by LLMs, they still retained authenticity. Therefore, it is required to extract and preserve human characteristics in reviews.

A review filter based on the findings needs to be developed to eliminate not only AI-generated reviews but also uninformative human-written reviews. For example, many

human-written reviews on Amazon consisted of brief statements such as "good product" or "it's nice". The result of this study provides the basis for exploring how to filter out both less useful human-written reviews and AI-generated ones.

6. Conclusions

Amazon product reviews have been flooded with AI-written ones in the past few years. Those reviews were neither authentic nor useful to potential buyers. A total of 1866 Amazon product reviews were classified as either AI-generated or human-written with a 99.25% F1 score using TF–IDF features and SVC in this study. TF–IDF was effective in classifying AI-generated product reviews due to the relatively contained scope of the dataset. By focusing on product-related content, the effectiveness of TF–IDF can be improved. The method is fast and resource-efficient in detecting and removing AI-generated product reviews. The findings of this study contribute to making the internet a more trustworthy space by promoting genuine, human-written reviews.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The 6217 Amazon product reviews dataset compiled by the author can be found here: https://github.com/JialuenYang/Detection-of-AI-Generated-Product-Reviews-on-Amazon (accessed on 15 October 2024).

Conflicts of Interest: The author declares no conflicts of interest.

References

1. CNBC. Available online: https://www.cnbc.com/2023/04/25/amazon-reviews-are-being-written-by-ai-chatbots.html (accessed on 21 August 2024).
2. About Amazon. Available online: https://www.aboutamazon.com/news/policy-news-views/how-ai-spots-fake-reviews-amazon (accessed on 24 August 2024).
3. Dugan, L.; Ippolito, D.; Kirubarajan, A.; Shi, S.; Callison-Burch, C. Real or Fake Text? Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 12763–12771. [CrossRef]
4. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text Be Reliably Detected? *arXiv* **2024**, arXiv:2303.11156. [CrossRef]
5. Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; Wu, Y. How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv* **2023**, arXiv:2301.07597. [CrossRef]
6. Tang, R.; Chuang, Y.-N.; Hu, X. The Science of Detecting LLM-Generated Text. *Commun. ACM* **2024**, *67*, 50–59. [CrossRef]
7. Bhattacharjee, A.; Liu, H. Fighting Fire with Fire: Can ChatGPT detect AI-generated text? *ACM SIGKDD Explor. Newsl.* **2024**, *25*, 14–21. [CrossRef]
8. Prova, N. Detecting AI Generated Text Based on NLP and Machine Learning Approaches. *arXiv* **2024**, arXiv:2404.10032. [CrossRef]
9. Salminen, J.; Kandpal, C.; Kamel, A.M.; Jung, S.-G.; Jansen, B.J. Creating and detecting fake reviews of online products. *J. Retail. Consum. Serv.* **2022**, *64*, 102771. [CrossRef]
10. Originality.ai. Available online: https://originality.ai/blog/amazon-ai-generated-reviews (accessed on 26 August 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.