

Detecting AI-Generated Reviews for Corporate Reputation Management

R. E. Loke^a and M. GC

Centre for Market Insights, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Keywords: Natural Language Processing (NLP), Generative Language Models, Machine Learning, Benchmark Datasets.


Abstract: In this paper, we provide a concrete case study to better steer corporate reputation management in the light of recent proliferation of AI-generated product reviews. Firstly, we provide a systematic methodology for generating high quality AI-generated text reviews using pre-existing human written reviews and present the GPTARD dataset for training AI-generated review detection systems. We also present a separate evaluation dataset called ARED that contains a sample of product reviews from Amazon along with their predicted authenticity from incumbent tools in the industry that enables comparative benchmarking of AI-generated review detection systems. Secondly, we provide a concise overview of current approaches in fake review detection and propose to apply an overall, integrated group of four predictive features in our machine learning systems. We demonstrate the efficacy of these features by providing a comparative study among four different types of classifiers in which our specific machine learning based AI-generated review detection system in the form of random forest prevails with an accuracy of 98.50% and a precision of 99.34% on ChatGPT generated reviews. Our highly performant system can in practice be used as a reliable tool in managing corporate reputation against AI-generated fake reviews. Finally, we provide an estimation of AI-generated reviews in a sample of products on Amazon.com that turns out to be almost 10%. To validate this estimation, we also provide a comparison with existing tools Fakespot and ReviewMeta. Such high prevalence of AI-generated reviews motivates future work in helping corporate reputation management by effectively fighting spam product reviews.

1 INTRODUCTION

In this digital age, corporate reputation management has become a critical aspect for businesses worldwide (Ramos & Casado-Molina, 2021). With the widespread use of social media and online platforms, companies face an increasingly complex challenge in maintaining their image and brand reputation (Killian & McManus, 2015). Online reviews and ratings have become a significant source of information for consumers, and they heavily influence purchasing decisions (Dwidienawati, Tjahjana, Abdinagoro, Gandasari, & Munawaroh, 2020). Over 80% of consumers in the United States rely on online reviews when making a purchase decision (Smith & Anderson, 2016). Therefore, it is essential for businesses to have reliable processes and systems to manage and monitor their online reputation by monitoring the reviews.

Positive reviews are crucial to a brand as well as a marketplace's reputation, as they can influence customer perception of both brands and the marketplace. Electronic Word-of-Mouth (eWOM), in the form of online product reviews, plays a significant role in shaping consumers' purchase choices (Salminen, Kandpal, Kamel, Jung, & Jansen, 2022). However, there are instances where certain individuals may attempt to damage a company's reputation by generating fake negative reviews and posting them online (Salehi-Esfahani & Ozturk, 2018).

There are two primary ways fake reviews are being generated and disseminated in a large scale: (i) purchasing human reviews using unethical incentives and (ii) using AI systems such as ChatGPT to generate plausible yet fake reviews (Salminen, Kandpal, Kamel, Jung, & Jansen, 2022). The human-oriented approach involves businesses paying

^a <https://orcid.org/0000-0002-7168-090X>

individuals or groups of people to write fake reviews manually in return of monetary or product incentives. In contrast, the AI-generated method involves the use of generative language models to automate the creation of human-like reviews. These algorithms can generate text that is nearly indistinguishable from genuine reviews, making it increasingly challenging to detect fake reviews at scale. As such, businesses must be vigilant and proactive in monitoring customer reviews and take steps to combat the spread of fake reviews to ensure that their online reputation remains intact and that their customers have access to accurate and trustworthy information.

With the recent surge of advancements in artificial intelligence, many organizations have started to adopt the AI-driven tools and technology to gain a competitive edge. AI has proven to be indispensable to analyze large volumes of data, including textual content such as online reviews, to provide valuable insights that can help businesses manage their reputation (Davenport & Rajeev, 2018). Powerful technologies like Large Language Models (LLMs) make it possible for individuals to generate fake reviews efficiently and effortlessly, leading to concerns about the accuracy and reliability of online reviews (Yao, Viswanath, Cryan, Zheng, & Zhao, 2017). Fake AI-generated reviews can be easily found in marketplaces like Amazon. This poses a significant threat to brands and marketplaces and thus it is prudent to take proactive measures to identify these as early as possible. Developing systems that can flag AI-generated reviews vs human written reviews can add one more layer to protecting corporate reputation.

Systems for detecting fake or malicious reviews in general are commonplace in ecommerce businesses. However, with recent proliferation and democratization of highly powerful AI systems like ChatGPT, systems able to detect AI-generated reviews are lacking. By developing and deploying AI-generated review detection systems, businesses can mitigate the potential harm caused by fake reviews and ensure that customers have access to accurate and trustworthy information.

In this paper, we propose an AI-generated review detection system for managing corporate reputation. Our objective is to develop a system that can automatically detect and classify online reviews. The proposed system is a machine learning classifier that uses natural language processing techniques to classify reviews as either genuine or AI-generated with a high accuracy.

The paper is organized in a logical manner beginning with the literature review (section 2), followed by the research methodology (section 3) and

the presentation of results (section 4). In section 5, discussion and limitations on the findings are thoroughly addressed.

2 LITERATURE REVIEW

This section provides a comprehensive review of corporate reputation and its relationship with online feedback. We then introduce generative language models, which is the technology that has enabled the recent surge of AI-generated reviews. Finally, we provide the review of deceptive and AI-generated review detection literature.

2.1 Corporate Reputation

Corporate reputation can be defined as the overall perception, evaluation, and assessment of a company or organization's character, actions, and performance in the eyes of its stakeholders, including customers, investors, employees, regulators, suppliers, and the general public. It represents the collective image and impression that the company has built over time through its behaviour, communication, products, and services. According to Crosier (1997), corporate reputation is an intangible asset that can influence a company's financial performance, risk management, and stakeholder engagement. A strong corporate reputation can create a positive perception of the company, resulting in increased customer loyalty, lower employee turnover rates, and enhanced brand recognition. Therefore, corporate reputation can be viewed as a value-creating strategy that contributes to a company's long-term success and sustainability and management of corporate reputation has emerged as a paramount factor in achieving business success.

Corporate reputations are built and maintained through a range of activities, including corporate social responsibility, communication strategies, product quality, and innovation (Crosier, 1997). Companies can also develop a strong brand identity that resonates with their target audience and aligns with their core values and mission (Gebhart, 1996). In recent decades, the activities and indicators of corporate reputation have largely shifted to the online medium, for example through social media marketing, online feedback, and ecommerce product reviews. Dellarocas et al., (2007) highlighted the value of online product reviews in forecasting sales and emphasized the need for companies to monitor and manage the authenticity of reviews to maintain their credibility and reputation. As such, companies can leverage online reputation management strategies

to monitor, respond, and mitigate online feedback that can impact their reputation (Proserpio & Zervas, 2017).

The quality and sentiment of online feedback can profoundly impact a company's reputation and result in the erosion of customer trust, loyalty and financial performance if not managed proactively. Therefore, it is of vital importance that companies advance the technologies that can actively monitor attempts of deception and degradation of their corporate reputation.

2.2 Generative Language Models

Generative language models are machine learning models that can generate human language as their outputs, conditioned on input text. Thanks to the advent of Transformer models which are a highly efficient and powerful neural network architecture, generative language models have seen impressive improvements (Vaswani, et al., 2023). With massively parameterized models trained on internet-scale data, Transformers have enabled generative language models that produce natural language outputs indistinguishable from human writing.

Transformer-based generative language models consist of an encoder-decoder architecture with self-attention mechanisms. The encoder processes the input text and learns to represent each word or token in the context of the entire sequence. The decoder generates the output text based on the encoded representations and the provided context (Vaswani, et al., 2023).

Training of a generative language model follows a two-stage process: pre-training and fine-tuning. During the pre-training stage, the model is trained on a large corpus of publicly available text from the internet in an unsupervised manner by learning to predict the next word in a sentence given the preceding words. The objective is to maximize the likelihood of the correct next word, which allows the model to learn the statistical patterns, concepts, and relationships within the text without supervised human labels. The pre-training process optimizes multiple layers of Transformers that may include millions or even billions of parameters.

Once the model has been pre-trained on a vast amount of text data, it is fine-tuned on specific downstream tasks. This involves training the model on a narrower dataset that is specific to the target application, with or without labelled annotations. For example, the model can be fine-tuned on a dataset of news articles or scientific papers to generate text in those domains. Fine-tuning allows the model to adapt

its knowledge to the specific requirements of the task and to produce more accurate and contextually relevant outputs.

Probabilistic sampling is used to produce coherent and diverse responses conditioned on an input text (Holtzman et al., 2020). The sampling process involves generating the next word or token based on the probabilities assigned by the model to each possible word in the vocabulary. These probabilities are determined by the model's knowledge synthesized during the training phase. The sampling methods generate the most likely next words given the current input, but also introduce some level of randomness to ensure diversity in the generated text. In this way, large language models with billions of parameters are trained on vast amount of textual data which can then be used to generate highly plausible text.

Services such as ChatGPT (<https://chat.openai.com>) and Anthropic Claude (<https://claude.ai>), launched on November 30, 2022 and March 14, 2023, respectively, have enabled access to extremely powerful language models. These services make access to generate language models extremely democratized for users with a plethora of possible applications.

2.3 Deceptive and AI-Generated Review Detection

The rise of deceptive fake reviews has emerged as a major concern for consumers and businesses alike (Lee, Song, Li, Lee, & Yang, 2022). Fake reviews can generally be defined as reviews that are deliberately created to deceive readers and manipulate ratings (Ott, Choi, Cardie, & Hancock, 2011).

Given the ease of access to powerful generative language models such as ChatGPT and Claude, AI-generated reviews are becoming increasingly prevalent and difficult to detect which is corroborated by a surge of research in this area. For example, Lee et al., (2022) recently used supervised machine-learning to detect fake reviews from statistical features of review text and found that machine-learning classifiers outperform traditional hand-crafted and rule-based approaches. In a different vein from using textual features, Salminen et al., (2022) presented a method to identify fake reviews using review ratings, review frequency, and reviewer credibility.

Several industrial tools and services have also attempted to tackle the problem of detecting fake reviews. For example, Thereviewindex (<https://thereviewindex.com>), Fakespot

(<https://www.fakespot.com>) and ReviewMeta (<https://reviewmeta.com>) are some of the most easily available tools to detect fake reviews (Basic, 2020; Chatfield, 2021). Thereviewindex uses machine learning techniques, specifically sentiment analysis, to identify patterns in the language used in reviews that suggest that they may be fake. Similarly, FakeSpot and ReviewMeta use a combination of NLP and statistical methods to detect fake reviews (Awad, Salameh, Ngoungoure, & Abdullah, 2022).

While most of the tools and research has been focused on human-generated fake reviews, the problem of AI-generated fake reviews is rapidly becoming highly important. The growing prevalence of AI-generated reviews is a concern for companies that rely on reviews to attract customers, as they can be difficult to detect and have the potential to significantly impact a company's reputation. Zhang & Ghorbani (2020) provided an overview of online fake news, including the problem of AI-generated fake reviews, and discussed the challenges associated with detecting and mitigating it. These authors also observed a lack of sufficiently large and effective benchmark datasets.

An early seminal work by Jindal & Liu., (2008) proposed a dataset consisting of 400 truthful reviews and 400 human-written deceptive reviews written by the help of Amazon Mechanical Turk workers. Recent works have focused on bootstrapping the machine generated reviews dataset from already available human written datasets using generative models. For example, Salminen et al. (2022) used GPT-2 to create a dataset for a classification task for fake review detection. Similarly, Shehnepoor et al. (2021) designed a system that generates bot reviews given a set of real reviews which consists of genuine reviews and fraud human reviews.

In our analysis, we categorize the AI-generated review detection literature into two categories based on the modelling technique as described below.

2.3.1 Deep Learning-Based Detection

An early prominent work in detecting AI-generated reviews was done by Elmurngi & Gherbi (2018) who proposed a system that leverages deep learning models to detect fake reviews and thereby enhances the accuracy of sentiment analysis for online reviews. Aghakhani et al. (2018) proposed to use a generative adversarial network (GAN) to generate fake reviews that mimic the language patterns and characteristics of real reviews which are then subsequently used to train a deep-learning model that detects fake reviews. Similarly, Shehnepoor et al. (2021) developed a deep

learning-based model that uses adversarial training to generate synthetic reviews and improve the detection of fake reviews. Mohawesh et al. (2021) performed a comparative study of many classifiers such as C-LSTM, HAN, Convolutional HAN, Char-level C-LSTM, BERT, DistilBERT and RoBERTa out of which RoBERTa performed the best with 91% accuracy. Similarly, Salminen et al. (2022) proposed fine-tuning the RoBERTa model for classifying AI-generated review texts and achieved an accuracy of 96%. Zhang et al. (2023) proposed an AI-generated review detection system that uses an attention mechanism and a convolutional neural network to extract features from texts in order to be able to classify reviews as genuine or fake.

2.3.2 Feature-Based Detection

An alternative approach to using deep learning models directly on textual content is the feature-based approach. It is grounded on the assumption that there are explicit human-defined dimensions in which human-written and machine-generated texts differ. High-level features, built upon the disparities between human and machine text, can provide a transparent and comprehensible approach to detecting AI-generated fake reviews. Additionally, such features offer valuable insights into the distinctive behaviour of language models making the detection process more explainable and interpretable (Badaskar, Agarwal, & Arora, 2008). An early prominent feature-based method for natural language was proposed by Argamon-Engelson et al. (1998) who employed extraction of stylistic features for text classification and introduced the concept of Stylometry, which has since been successfully applied in various tasks. Desaire et al. (2023) recently proposed a model using the feature-based approach and reported an impressive 99% accuracy on ChatGPT generated text detection.

3 METHODOLOGY

This section provides a concise overview of the methodology used in our study, including the dataset generation, feature selection and the specific machine learning methods employed.

3.1 Dataset Generation

We choose Amazon reviews as our subject of study as (Wood, 2023) performed a thorough analysis of 720 million Amazon reviews and discovered that

approximately 42% of them were determined to be fraudulent. Such findings highlight the alarming prevalence of fake reviews in the ecommerce industry, particularly on a marketplace like Amazon. This makes Amazon reviews an ideal testing case for a systematic study on AI-generated reviews through the lens of corporate reputation management.

For our experimental design, we create two datasets. GPT Amazon Reviews Dataset (GPTARD): A dataset composed of 1500 human-written and 1500 ChatGPT generated Amazon reviews for training and evaluation of classification models. Amazon Reviews Evaluation Dataset (ARED): A dataset composed of 1200 reviews across 128 Amazon products posted after release of ChatGPT (November 30, 2022), along with obtained results of fake-review detection using incumbent tools FakeSpot and ReviewMeta.

We describe the methodology and motivation for the generation of these datasets below.

3.1.1 GPT Amazon Reviews Dataset (GPTARD)

In the absence of an established dataset for AI-generated detection, we follow the direction of Salminen et al. (2022) and use pre-existing human written reviews to bootstrap our own AI-generated reviews for the purpose of training and evaluating our models as described below. We base this dataset on the Stanford Network Analysis Project (SNAP) (Leskovec & McAuley, 2013) that contains 34 million Amazon reviews on 2 million products, collected in 2014.

Initially, we sample a subset of 5000 datapoints from the SNAP dataset. We only include reviews with a word count greater than the 25th percentile of the full dataset, specifically, a count of 33, to ensure that the sample has higher quality, longer reviews. Additionally, to ensure that our sample consists of a diverse set of products, since no product category labels are available in the SNAP dataset, we perform topic modelling on the content of the reviews and down sample to 3000 data points so that no single topic comprises in total more than 6% of the sample. This ensures that our dataset is not biased towards a particular type of product. The topic modelling was done by using TF-IDF methodology as outlined in Maarten (2022).

Next, we sample 1500 data points from the 3000 as candidates to bootstrap AI-generated reviews. For each of the candidates, we extract three pieces of information from the original human-written review to generate a synthetic review using OpenAI ChatGPT API, specifically the “gpt-3.5-turbo”

model: (1) *Prefix*---The initial 20 words of the review which remain unchanged in the generated review to prevent the generation from going off-topic. (2) *Polarity*---The sentiment polarity, either “positive” or “negative”. This is provided in the original SNAP dataset. (3) *Max Length*---To prevent a length bias in downstream classification, we constraint the generations to be the same length as the original review.

Finally, we prompt the API to complete generation in the following format: “Complete the following {*Polarity*} review of {*Max Length*} words: {*Prefix*}”. The data generation process was carried out on May 28, 2023, and it took approximately 30 minutes to generate all 1500 reviews. The API usage incurred the cost of 8 euros. Thus, we created a dataset of 1500 genuine human-written reviews and 1500 ChatGPT generated reviews for our experiments.

3.1.2 Amazon Reviews Evaluation Dataset (ARED)

In addition to GPTARD dataset for training and evaluating models, we created an additional dataset that we refer to as Amazon Reviews Evaluation Dataset (ARED). The goal of this dataset is to provide a more recent collection of Amazon reviews as opposed to GPTARD that facilitates comparison and evaluation of our proposed AI-generated review detection system with incumbent tools in the industry, namely FakeSpot (FakeSpot, sd) and ReviewMeta (ReviewMeta, sd).

We implemented a custom scraping tool in Python based on Scrapy (Kouzis-Loukas, 2016) to scrape Amazon reviews published after November 30, 2022, aligning with the release date of ChatGPT (Dhedra, 2023). It is essential to consider this timeframe, as ChatGPT-generated text can only have been written after that date. Importantly, we reference to the study of He et al. (2022) who determined the 13 top product categories associated with fake reviews on Amazon. We randomly selected 12 products within these categories, to scrape, on July 14, 2023, 100 product reviews that were dated in the interval from December 2022 to July 2023.

Finally, for each of the scraped reviews, we applied the public available FakeSpot and ReviewMeta tools. This allows us to determine and benchmark the proportion of fake reviews in the dataset.

Thus, with ARED, we created a specific evaluation dataset that is separate from GPTARD to estimate the performance of our system in a real

world setting by comparing its performance with established industrial tools.

3.2 Feature Selection

In the development of our machine learning based AI-generated text detection models, we opt to use hand-crafted features guided by previous works done in generative language models as discussed in section 2.3. Specifically, we adopt the repetitiveness and diversity measures from Su et al. (2022), the perplexity measure as used by Holtzman et al. (2020) and the part-of-speech tags as supported by Clark et al. (2019). Although these four feature types completely cover the well-known hand-crafted feature category, it is nevertheless the first time that all types are combined in a single study. Since we do content-based analysis, all our selected features are derived solely from the text content of the reviews and not from any associated metadata such as reviewer's profile, date of posting etc.

In this paper, we hypothesize and stipulate that all these features have discriminative properties for detection of synthetic texts generated by generative language models that should consequently be used all together to feed classifiers.

3.2.1 Repetitiveness

Repetitiveness refers to using a subset of the vocabulary disproportionately and it has been identified as a key property of generative language by several works in the literature. According to Holtzman et al. (2020), language models tend to rely on frequent words in machine-generated texts and thus result in excessive repetition and a lack of diversity. Similarly, Ippolito et al. (2020) found that approximately 80% of the probability mass in machine-generated language is concentrated in the 500 most common words. Additionally, Holtzman et al. (2020) also highlighted the low variance of next-token probabilities in machine-generated text, indicating a lack of exploration into low-probability zones as observed in human text. Gehrmann et al. (2019) identified the prevalent issue of highly parallel sentence structures in machine-generated text, while Jiang et al. (2020) noted occasional repetition of entire phrases.

Repetitiveness formulation as proposed by (Su, et al., 2022) assesses sequence-level repetition in the generated text by measuring the extent to which duplicate n -grams occur within it. An n -gram is a subsequence of consecutive words, where the value of " n " represents the number of words in each

subsequence. For example, a 2-gram, also known as a bigram, considers of two adjacent words, while a 3-gram or trigram considers three consecutive words. This metric measures the proportion of duplicate n -grams in the generated text, indicating the level of repetition within a text segment.

In our implementation, we convert all the text tokens to lowercase and filter out the stop words using NLTK library Bird et al. (2009) ensuring the metric is focused on meaningful words. Following Su et al. (2022), we extract the bigrams, trigrams, and 4-grams from the tokenized sentences. Finally, we calculate the repetition scores for each n -gram category following Equation 1 that scores repetitiveness at n -gram level.

$$\begin{aligned} rep_n \\ = 100 \times \left(1.0 - \frac{|unique\ n - grams(\hat{x})|}{|total\ n - grams(\hat{x})|} \right) \end{aligned} \quad (1)$$

3.2.2 Diversity

Syntactic diversity of text refers to the repetition of similar syntactical structure throughout the text segment. Gehrmann et al. (2019) and Zellers et al. (2020) shed light on the issue of syntactic diversity in language models. They highlight a models' tendency to rely on repetitive expressions and a consequent lack of syntactic and lexical diversity. Such models can fail to utilize synonyms and references in the same way as humans, resulting in limited variation in machine-generated text. See et al. (2019) find that the generated texts contain a higher proportion of verbs and pronouns, while nouns, adjectives, and proper nouns are relatively scarce. This discrepancy in syntactic distribution usage can be a useful characteristic of machine-generated text for detection.

To calculate the diversity, we use summation-based formulation in Equation 2 as proposed by Su et al. (2022). Here, rep_n is given by Equation 1.

In our implementation, we utilize the computed repetitiveness scores per n -gram to derive the overall diversity score using Equation 2. For example, if we have the repetitiveness scores 8, 4 and 2 respectively for bigrams, 3-grams and 4-grams, the diversity score is computed to be 2.86.

$$Diversity = \sum_{n=2}^4 \left(1.0 - \frac{rep_n}{100} \right) \quad (2)$$

3.2.3 Part of Speech (PoS)

Part of speech (PoS) tagging is a fundamental task in natural language processing (NLP) that involves assigning the appropriate PoS labels, for example noun, pronouns, adverbs, to words in a sentence to represent the syntactic structure and meaning of textual data (Manning et al., 2014). PoS distribution has been recognized as significant in distinguishing between human and machine-generated texts (Clark et al., 2019).

Clark et al. (2019) highlight the significance of PoS distribution in distinguishing between human and machine-generated texts. As proposed by Feng et al. (2010), we use the review-level count of nouns and use the NLTK library (Bird, Klein, & Loper, 2009) to extract (NOUN), pronouns (PRON), verbs (VERB), adjectives (ADJ), adverbs (ADV), determiner (DET), conjunction (CONJ), numeral (NUM), and particle (PRT).

3.2.4 Perplexity

In language modelling literature, perplexity score gauges the level of uncertainty or surprise in predicting the following word in a sequence, considering the preceding words (Tang, Chuang, & Hu, 2023). It is computed by averaging the negative average log-likelihood of the language model on the given text or dataset. Log-likelihood measures how likely the predicted word is based on the model's internal probabilities. By taking the average log-likelihood and negating it, it can be converted into the perplexity score that indicates uncertainty or surprise rather than likelihood. A lower perplexity score indicates more aligned predictions to model's knowledge and understanding of the underlying language patterns.

Perplexity serves as a crucial metric for evaluating generated text in language generation research. Holtzman et al. (2020) compared various decoding strategies and their perplexity scores against a reference text. The authors emphasized that excessively low perplexity frequently leads to the production of repetitive and less diverse output, reminiscent of the characteristics commonly associated with computer-generated text. Similarly, more studies have shown that language models have a propensity to focus on prevalent patterns found in their training texts, resulting in low perplexity scores for the text they generate (Tang, Chuang, & Hu, 2023; Fu, Lam, So, & Shi, 2021).

Given a tokenized input sequence $X = (x_0, x_1, \dots, x_t)$, the perplexity of X can be defined as depicted in Equation 3 (huggingface, n.d.). In this context, the

expression $\log p\theta(\hat{x}_i|x_{<i})$ represents the log-likelihood of the i^{th} token given the preceding tokens $x_{<i}$ as predicted by our model. For example, given the phrase "The cat", a language model predicts the next word to be "is" with a certain probability. The log-likelihood $\log p\theta("is"|"The cat")$ represents the logarithm of the likelihood assigned by the model to the word "is" given the preceding the words "The cat".

$$PPL(X) = \exp \left\{ -\frac{1}{t} \sum_i^t \log p\theta(\hat{x}_i|x_{<i}) \right\} \quad (3)$$

To compute perplexity, we begin by dividing longer texts into smaller chunks and iteratively calculating the negative log-likelihood (NLL) loss of the GPT-2 Radford et al. (2018) model using the Hugging Face Transformers library (Wolf, Lysandre, Victor, Julien & Moi, 2020). Finally, The NLL values for each chunk are collected, and the average NLL are exponentiated to obtain the perplexity value for each review. A lower perplexity indicates better prediction performance and higher model confidence.

3.3 Classification Models

Our proposed approach for AI-generated review detection is based on hand-crafted features from natural language. This enables us to experiment with a widely used range of machine learning models in AI-generated review detection literature from which we pick four models. (1) Logistic regression is a probabilistic classification algorithm that models the relationship between a set of input features and the binary outcome using the logistic function (Bishop, 2006). (2) Random Forest is an ensemble learning algorithm that combines multiple decision trees to perform classification or regression tasks (Paul et al., 2018). Random Forests are known for their robustness, scalability, and ability to handle high-dimensional data, making them popular for various machine learning tasks. The combination of multiple trees helps to reduce overfitting and improve the overall generalization performance of the model. (3) Gradient Boosted Random Forest, of which the most popular implementation is XGBoost (Chen et al., 2016), is a powerful and widely used ensemble learning algorithm that combines the strengths of both Gradient Boosting and Random Forest. With its ability to handle missing values, handle both categorical and numerical features, and exploit parallel computing, XGBoost has gained popularity

and achieved state-of-the-art performance in various machine learning competitions and real-world applications. (4) Artificial neural networks are computational models inspired by the structure and functioning of the human brain (Geiger, 2021). By leveraging nonlinear activation functions and multiple hidden layers, deep neural networks can model complex relationships in data, making them capable of handling a wide range of tasks, including highly challenging domains such as computer vision and natural language processing.

4 EXPERIMENTS AND RESULT ANALYSIS

4.1 Experimental Details

4.1.1 Datasets

We perform our experiments on two datasets, namely the GPT Amazon Reviews Dataset (GPTARD) for training and validating our classification models, and Amazon Reviews Evaluation Dataset (ARED) for testing the effectiveness of our approach in identifying AI-generated online reviews. The details of the procedure for generating these datasets are described in section 3.3.

In Table 1, we provide the summary statistics of the two datasets used in our experiments and analysis.

Table 1: Summary statistics of the datasets used in the experiments.

Dataset	GPTARD	ARED
Data Source	SNAP Dataset + ChatGPT API	Scraped Amazon user reviews (11/22-06/23)
Mean review length	143.76	59.35
Std review length	69.54	71.96
# Training	2400	-
# Test	600	1200

4.1.2 Implementation and Hyperparameter Details

Our codebase is based on Python and popular machine learning libraries. Specifically, we use Scikit-Learn for implementing Logistic Regression, Neural Networks and Random Forest models. We also use XGBoost, a popular and highly efficient gradient-boosted tree model. Finally, we use NLTK for regular NLP tasks such as Part of Speech

recognition and n-gram extraction. Our data and code is publicly available in a repository at <https://github.com/MasterDDB22/mpMallika>.

In Table 2, we provide the hyperparameter details for different models used in our experiments. Unless specified in the table below, we use the default hyperparameters in the implementations provided in the libraries.

Table 2: Hyperparameters used for different models.

Model	Hyperparameter	Value
Logistic Regression	Penalty	L2
Neural Network	Hidden Layer Sizes	(100)
Neural Network	Learning Rate	0.001
Neural Network	Solver	Adam
XGBoost	Max Depth	6
XGBoost	Number of Estimators	100
Random Forest	Number of Estimators	100

4.2 Classification Performance Analysis

First and foremost, we are interested in the performance of several machine learning classifiers in the binary classification of GPT Amazon Reviews Dataset (GPTARD). To find the best performing classification model, we experiment with Logistic Regression (LR), Multilayer Perceptron Neural Network (NN), Random Forest (RF) and XGBoost gradient-boosted tree models as described in section 3.3.

We report the mean accuracy metric on the test split of the dataset for 5 different runs of training, along with the standard deviation in Table 3. We can observe that the Random Forest model outperforms all the other classifiers in test set accuracy.

However, it is also clear that there is no major difference in the performance of the other classifiers. Given that the simpler linear logistic regression model also performs relatively well on the test data, we can conclude that the features used as the input to these models are highly discriminative in detection of AI-generated product reviews.

Additionally, we also report the precision, recall and F1 score metrics on the performance of the Random Forest model in Table 4.

We can observe that the precision of the model on detecting ChatGPT generated reviews is high, but the recall is a bit lower. This is a desirable property of a fake review detection system where we do not want to label genuine human written reviews as AI-generated. It is also interesting to note that the model has similarly high performance on recall for human written reviews, albeit with a trade-off on the

precision.

Table 3: Mean accuracy and standard deviation of accuracies for 5 runs on GPTARD test split.

Classification Model	Mean Accuracy (%)	Std Dev Accuracy (%)
Logistic Regression (LR)	97.66	1.11×10^{-16}
Neural Network (NN)	97.83	0.003
XGBoost	98.33	1.11×10^{-16}
Random Forest (RF)	98.50	0.002

Table 4: Precision, Recall and F1 score on GPTARD test split for best performing model, i.e., Random Forest.

	Precision (%)	Recall (%)	F1-score (%)
Human	96.95	99.31	98.11
ChatGPT	99.34	97.12	98.22

4.3 Evaluation on Recent Amazon Reviews

We are also interested in how our system performs in a real-world scenario. To this end, we utilize our Amazon Reviews Evaluation Dataset (ARED) which comprises of 1200 reviews across various products collected after the release of ChatGPT. Along with the reviews, this dataset also consists of estimations of fake reviews for each product by popular tools FakeSpot and ReviewMeta as described in section 3.1.2.

Since we do not have ground truth labels for which of these reviews are AI-generated, we report prevalence in Table 5 which is the total percentage of reviews predicted as AI-generated by our system in the entire dataset. We can observe that our best performing classifier, Random Forest model, estimates that around 9.82% of the 1200 reviews in the ARED dataset are AI-generated. Other classifiers also report similar prevalence, with lowest number being around 2% lower reported by the neural network model.

To provide a comparison, we also report the average prevalence of deceptive reviews as estimated by ReviewMeta and FakeSpot tools. It can be observed that ReviewMeta and FakeSpot report 22.86% and 33.06%, respectively, which includes all different forms of “deceptive” reviews such as incentivized reviews, AI-generated reviews, and human-written fake reviews. Therefore, our system which is trained specifically to detect AI-generated reviews is a subset of the other two providers and thus reports a lower prevalence. This is in line with the expectation of how it would perform in a real-world situation.

Table 5: Prevalence of AI-generated reviews in ARED dataset for different models.

Classification Model	Prevalence (%)
Logistic Regression (LR)	8.65
Neural Network (NN)	7.62
XGBoost	9.23
Random Forest (RF)	9.82

5 DISCUSSION

We performed a thorough review of fake-review detection literature and concluded that while there has been a significant amount of work on human-written deceptive reviews, there is still room for research in AI-generated reviews, given the recent democratization of tools like ChatGPT. Our work was motivated to fill this gap and we proposed several novel methodologies to help develop a detection system for AI-generated reviews. By comparing the prevalence of such reviews in a sample of Amazon.com product listings, we discovered that around 10% of the products in the sample had reviews generated by language models. We verified this finding by comparing the prevalence with established tools like FakeSpot and ReviewMeta. Thus, we conclude that the issue of AI-generated product reviews is significant from the perspective of corporate reputation as such reviews are easy to generate in mass and have the potential to hurt a company’s brand as well as finances.

We identified two common approaches, namely, deep learning-based and feature-based detection, in the literature. While deep learning-based approaches are motivated by the widespread success of such models in a variety of different areas in machine learning, we were motivated by recent success, flexibility, and interpretability of feature-based methods. We validated the effectiveness of such feature-based methods by developing a machine learning based system with a precision of 99.34% in detecting ChatGPT written reviews.

We identified a lack of benchmark datasets in the AI-generated review detection research and proposed a systematic way of using pre-existing human-written reviews to generate high quality and large amount of AI-generated review using OpenAI ChatGPT API. This dataset, that we refer to as GPTARD, was used to train and evaluate different machine learning classifiers in our work. We also realized that it is of paramount importance that we can evaluate performance of such systems in a real-world scenario and should facilitate comparative analysis of performance. To this end we created another bespoke

dataset, that we refer to as ARED, to help evaluate proposed systems by comparing them to the performance of industry standard tools.

Based on recent works on improving the generation quality of language models, we identified four features having highly discriminative properties, namely, Perplexity, Diversity, Repetitiveness and Part of Speech distribution of generated text. We performed an additional study on efficacy of these features compared to other commonly used features such as statistical measures and readability scores of generated texts and demonstrated the superiority of the features we identified, nevertheless, due to space requirements, we had to omit this section from the current version of this paper.

We evaluated several machine learning models and identified that Random Forest model trained on the four feature groups we identified can achieve 98.50% accuracy on proposed GPTARD dataset, with 99.34% precision in detection of reviews written by ChatGPT language model. The benefits of our model are that given its feature-based nature, results are relatively easy interpretable, and processing does not demand high computational requirements to deploy, unlike deep learning-based models. This high performance and practicality of our system is encouraging and permits that AI-generated reviews can be detected reliably and efficiently. Therefore, we strongly recommend businesses to adopt our current approach to be able to combat AI-generated reviews.

5.1 Limitations

We find that with thoughtful dataset generation, feature selection and evaluation, businesses can tackle the problem of AI-generated review detection in a practical and reliable manner. Our study proposed two datasets for training and evaluation of such detection systems and demonstrated their usefulness in developing AI-generated review detection. However, our datasets are based on just Amazon reviews. To provide a better source distribution of data that might even better represent a business's reputation dynamics, we could also incorporate data based on other sources, for example: TripAdvisor, Trustpilot etc.

Similarly, our datasets could be constructed by systematically representing rating and sentiment levels of reviews to study potential bias in the reported performance. Future research could also be directed to provide more detailed insights on product and product category levels.

Finally, given the widespread success of deep learning-based models and their ability to learn

powerful representations, more work could be beneficial in exploring deep learning models to provide a comparative analysis of benefits and drawback of feature-based and deep learning-based approaches, respectively.

We hope that these insights can be beneficial for businesses and researchers alike for developing practical systems for tackling the rising issue of AI-generated fake reviews.

ACKNOWLEDGEMENTS

This paper has been inspired on the MSc project of Mallika GC who was involved via the master Digital Driven Business at HvA. Thanks go to Stephanie van de Sanden as well as several anonymous reviewers for providing some useful suggestions to an initial version of this manuscript. Rob Loke is assistant professor data science at CMIHvA.

REFERENCES

- Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., & Vigna, G. (2018, May 25). Detecting Deceptive Reviews using Generative Adversarial Networks. <https://arxiv.org/pdf/1805.10364.pdf>
- Argamon-Engelson, S., Koppel, M., & Avneri, G. (1998). Style-based Text Categorization: What Newspaper Am I Reading? Proc. of the AAAI Workshop on Text Categorization, 1-4.
- Awad, M., Salameh, K., Ngoungoure, A.M., & Abdullah, M. (2022). Opinion Spamming: Analyzing the Accuracy of Online Detection Tools. CEEeGov '22: Proc. of the Central and Eastern European eDem and eGov Day, 142-146. New York: ACM.
- Badaskar, S., Agarwal, S., & Arora, S. (2008). Identifying Real or Fake Articles: Towards better Language Modeling. Proc. of the 3rd Int. Joint Conf. on Natural Language Processing: Volume-II. <https://aclanthology.org/I08-2115>
- Basic, M. (2020). Comparing Fake Review Tools on Amazon.com.
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.
- Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.
- Brown, T.B., Mann, B., Ryder, N., & Subbiah, M. (2020, July 22). Language Models are Few-Shot Learners. <https://arxiv.org/pdf/2005.14165.pdf>
- Chatfield, C. (2021, August). These 3 Tools Will Help You Spot Fake Amazon Reviews. <https://www.makeuseof.com/fake-reviews-amazon/>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proc. of the 22nd ACM SIGKDD Int.

- Conf. on Knowledge Discovery and Data Mining, 785-794.
- Clark, J., Radford, A. & Wu, J. (2019). <https://github.com/openai/gpt-2-output-dataset/blob/master/detection.md>
- Crosier, K. (1997). Corporate Reputations: Strategies for Developing the Corporate Brand. *European J. of Marketing*, 31(5-6). Corporate Reputations: Strategies for Developing the Corporate Brand. - Document - Gale Academic OneFile
- Davenport, T.H., & Ronanki, R. (2018). Artificial Intelligence for the Real World. *Harvard Business Review*, 96(1), 108-116. 3 Things AI Can Already Do for Your Company (hbr.org)
- Dellarocas, C., Zhang, X., & Awad, N.F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. of Interactive Marketing*, 21(4), 23-45.
- Desaire, H., Chua, A.E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science* 4, 101426.
- Dheda, G. (2023, June). When Was ChatGPT Released? <https://openaimaster.com/when-was-chatgpt-released>
- Dwidienawati, D., Tjahjana, D., Abdinagoro, S., Gandasari, D., & Munawaroh. (2020, Nov.). Customer review or influencer endorsement: which one influences purchase intention more? *Heliyon*, 6(11).
- Elmurngi, E., & Gherbi, A. (2018). Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques. *DATA ANALYTICS 2017: The 6th Int. Conf. on Data Analytics*
- FakeSpot. (n.d.). FakeSpot - Use AI to detect fake reviews and scams. <https://www.fakespot.com/>
- Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. *COLING 2010, 23rd Int. Conf. on Computational Linguistics*, Beijing.
- Fu, Z., Lam, W., So, A. M.-C., & Shi, B. (2021, March 22). A Theoretical Analysis of the Repetition Problem in Text Generation. <https://arxiv.org/pdf/2012.14660.pdf>
- Gebhart, J. (1996). Reputation: Realizing Value from the Corporate Image. *Sloan Management Review*, Cambridge, 37(2), 116.
- Gehrmann, S., Strobelt, H., & Alexander, R.M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111-116.
- Geiger, B. C. (2021). On Information Plane Analyses of Neural Network Classifiers—A Review. *IEEE Trans. on Neural Networks and Learning Systems*, 33(2), 7039-7051.
- He, S., Hollenbeck, B., & Proserpio, D. (2022). The Market for Fake Reviews. *Marketing Science*, 41(5), 896-921.
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020, February 14). The Curious Case of Neural Text Degeneration. *The Int. Conf. on Learning Representations (ICLR)*. <https://arxiv.org/pdf/1904.09751.pdf>
- huggingface. (n.d.). Perplexity of fixed-length models. <https://huggingface.co/docs/transformers/perplexity>
- Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic Detection of Generated Text is Easiest when Humans are Fooled. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 1808-1822.
- Jiang, S., Wolf, T., Monz, C., & Rijke, M. d. (2020, April 9). TLDR: Token Loss Dynamic Reweighting for Reducing Repetitive Utterance Generation. <https://arxiv.org/pdf/2003.11963.pdf>
- Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *WSDM '08: Proc. of the 2008 Int. Conf. on Web Search and Data Mining*. New York: ACM. <https://dl.acm.org/doi/abs/10.1145/1341531.1341560>
- Khalifah, S. (2021, September 16). The Truth Behind the Stars. <https://www.fakespot.com/post/the-truth-behind-the-stars>
- Killian, G., & McManus, K. (2015). A marketing communications approach for the digital era: Managerial guidelines for social media integration. *Business Horizons*, 58(5), 539-549.
- Koppel, M., Argamon, S., & Shimoni, A.R. (2022). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 401-412.
- Kouzis-Loukas, D. (2016). *Learning Scrapy*. Packt Publishing Ltd.
- Lee, M., Song, Y., Li, L., Lee, K., & Yang, S.-B. (2022). Detecting fake reviews with supervised machine learning algorithms. *Service Industries J.*, 1101-1121.
- Leskovec, J., & McAuley, J. (2013). Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. *RecSys*. <https://snap.stanford.edu/data/web-Amazon.html>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020, October 29). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880.
- Libai, B., Bart, Y., Gensler, S., Hofacker, C.F., Kaplan, A., Kösterheinrich, K., & Kroll, E.B. (2022). Brave New World? On AI and the Management of Customer Relationships. *J. of Interact. Marketing*, 51(1), 44-56.
- Mohawesh, R., Xu, S., Tran, S.N., Ollington, R., & Springer, M. (2021, April). Fake Reviews Detection: A Survey. *IEEE Access*, 9, 65771-65802.
- Maarten, G. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. <https://arxiv.org/abs/2203.05794>
- Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60, Baltimore.
- Oh, S. (2022). Predictive case-based feature importance and interaction. *Information Sciences*, 155-176.
- Ott, M., Choi, Y., Cardie, C., & Hancock, J.T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the

- Imagination. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 309–319, Oregon. <https://arxiv.org/abs/1107.4557v1>
- Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R., & Kundu, S. (2018). Improved Random Forest for Classification. *IEEE Trans. on Image Processing*, 27(8), 4012–4024.
- Perez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic Detection of Fake News. Proc. of the 27th Int. Conf. on Computational Linguistics, 3391–3401, Santa Fe, New Mexico.
- Proserpio, D., & Zervas, G. (2017). Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews. *Marketing Science*, 36(5), 645–665.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Raffel, C., Shazeer, N., Roberts, A., & Lee, K. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. of Machine Learning Research*, 1–67.
- Ramos, C.M., & Casado-Molina, A.-M. (2021, January). Online corporate reputation: A panel data approach and a reputation index proposal applied to the banking sector. *J. of Business Research*, 122, 121–130.
- ReviewMeta. (n.d.). ReviewMeta analyzes Amazon product reviews and filters out reviews that our algorithm detects may be unnatural. <https://reviewmeta.com/>
- Rubin, V. L., Conroy, N.J., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *eAssociation for Computational Linguistics: Human Language Technologies (NAACL-CADD2016)*.
- Smith, A., & Anderson, M. (2016, December 19). Online reviews. Retrieved April 18, 2023, from Pew Research Center: <https://www.pewresearch.org/internet/2016/12/19/online-reviews/>
- Salehi-Esfahani, S., & Ozturk, A.B. (2018). Negative reviews: Formation, spread, and halt of opportunistic behavior. *Int. J. of Hospitality Mngment*, 74, 138–146.
- Salminen, J., Kandpal, C., Kamel, A.M., Jung, S.-g., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *J. of Retailing and Consumer Services*, 64.
- See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C.D. (2019). Do Massively Pretrained Language Models Make Better Storytellers? Proc. of the 23rd Conf. on Computational Natural Language Learning, 843–861, Hong Kong, Assoc. for Comput. Linguistics.
- Shehnepoor, S., Togneri, R., Liu, W., & Bennamoun, M. (2021). ScoreGAN: A Fraud Review Detector based on Multi Task Learning of Regulated GAN with Data Augmentation. <https://arxiv.org/pdf/2006.06561.pdf>
- Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., & Collier, N. (2022, September 28). A Contrastive Framework for Neural Text Generation. <https://arxiv.org/pdf/2202.06417.pdf>
- Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review*, 20(3), 571–610. <https://www.jstor.org/stable/258788>
- Tang, R., Chuang, Y.-N., & Hu, X. (2023, June 2). The Science of Detecting LLM-Generated Texts. <https://arxiv.org/pdf/2303.07205.pdf>
- Touvron, H., Lavril, T., & Izacard, G. (2023, February 27). LLaMA: Open and Efficient Foundation Language Models. <https://arxiv.org/pdf/2302.13971.pdf>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2023, August 2). Attention Is All You Need. <https://arxiv.org/pdf/1706.03762.pdf>
- Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2019, September 26). Neural Text Generation with Unlikelihood Training. <https://arxiv.org/pdf/1908.04319.pdf>
- Wolf, T., Lysandre, S., Victor, C., Julien, D., & Moi, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45, ACM.
- Wood, R. (2023, May). Question of the Day: What percentage of Amazon reviews are potentially fake? <https://www.ngpf.org/blog/question-of-the-day/qod-what-percent-of-reviews-posted-on-popular-e-commerce-sites-are-fake/>
- Yao, Y., Viswanath, B., Cryan, J., Zheng, H., & Zhao, B.Y. (2017, September 8). Automated Crowdturfing Attacks and Defenses in Online Review Systems. <https://arxiv.org/pdf/1708.08151.pdf>
- Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2020, December 11). Defending Against Neural Fake News. <https://arxiv.org/pdf/1905.12616.pdf>
- Zhang, D., Li, W., Niu, B., & Wu, C. (2023). A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166, 113911.
- Zhang, X., & Ghorbani, A.A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102025.