Dissecting AI-Generated Fake Reviews: Detection and Analysis of GPT-Based Restaurant Reviews on Social Media

Alessandro Gambetti
Nova School of Business and Economics, alessandro.gambetti@novasbe.pt

Qiwei Han
Nova School of Business and Economics, qiweih@alumni.cmu.edu

Dissecting AI-Generated Fake Reviews: Detection and Analysis of GPT-Based Restaurant Reviews on Social Media

Completed Research Paper

Alessandro Gambetti
Nova School of Business and Economics
Carcavelos, Portugal
gambetti.alessandro@novasbe.pt

Qiwei Han
Nova School of Business and Economics
Carcavelos, Portugal
qiwei.han@novasbe.pt

Abstract

Recent advances in generative models such as GPT may be used to fabricate indistinguishable fake customer reviews at a much lower cost, posing challenges for social media platforms to detect this kind of content. This study addresses two research questions: (1) the effective detection of AI-generated restaurant reviews generated from high-quality elite authentic reviews, and (2) the comparison of out-of-sample predicted AI-generated reviews and authentic reviews across multiple dimensions of review, user, restaurant, and content characteristics. We fine-tuned a GPT text detector to predict fake reviews, significantly outperforming existing solutions. We applied the model to predict non-elite reviews that already passed the Yelp filtering system, revealing that AI-generated reviews typically score higher ratings, users posting such content have less established Yelp reputations and AI-generated reviews are more comprehensible and less linguistically complex than human-generated reviews. Notably, machine-generated reviews are more prevalent in low-traffic restaurants in terms of customer visits.

Keywords: *AI-generated Content, Natural Language Generation, Fake Review Detection, GPT, Social Media*

Introduction

Online reviews have served as valuable signals about product quality that may bridge information asymmetries between customers and sellers in online marketplaces, which in turn influence customer purchases (Duan et al. 2008; Hu et al. 2008; Ott et al. 2012; Vana and Lambrecht 2021). Fake reviews can be defined as opinion-based disinformation, which is fabricated and propagated by spammers with the ambition of misleading and deceiving customers (Paul and Nikolaev 2021). With the prevalence of review systems embedded in social media, such as TripAdvisor, Yelp, and Facebook, fake reviews have also been witnessed to proliferate on these platforms, aiming to mislead customers by pretending to be authentic, in order to achieve unjustly competitive gains for certain businesses (Jindal and Liu 2008; Ott et al. 2012; Ott et al. 2011). The COVID-19 pandemic may further exacerbate the issue because many less-experienced customers are forced to make more online purchases and tend to rely on reviews more heavily (McCluskey 2022). According to a report from World Economic Forum, fake reviews' economic impact on global online spending has reached $152 billion in recent years (Marciano 2021).

Given that the fundamental value of reviews is rooted in their authenticity that reflects customers' truthful experience, fake reviews would not only harm customers but also severely threaten to erode trust in online review systems and damage the reputation of social media platforms (He et al. 2022; Ma and Lee 2014). Typically, He et al. (2022) showed that there exists a market of fake reviews, and sellers choose to rely on paid review farms for content creation. To this end, major platforms propose countermeasures to fight review fraud with both manual analyses by the content moderation team and automated systems. For

example, since 2019, TripAdvisor has started to publish the transparency report outlining its effort to keep fake reviews off the site. The most recent report in 2021 revealed that over 2 million (3.6%) of reviews were determined to be fraudulent (TripAdvisor 2021). In a similar vein, Yelp implemented an automated recommendation software to filter off 4.3 million suspicious reviews out of 19.6 million reviews and has been displaying the most reliable and helpful reviews on the main business pages (McCluskey 2022).

However, social media still face significant challenges to counteract not only *user-generated* fake reviews but also *machine-generated* fake reviews, because of advances in generative large language models (LLMs) such as GPT-based models (OpenAI 2023), including ChatGPT, a chatbot that leverages their architecture to engage in human-like conversations and provide support to users in question-answering tasks. LLMs possess the ability to produce textual content that emulates human writing styles to the point where it becomes nearly indistinguishable from the human-generated text. On the one hand, fake reviews fabricated by LLMs trained on real reviews have become essentially indiscernible. One study employing currently obsolete AI text generators showed that machine-generated fake reviews could evade human detection and even receive a higher score of perceived usefulness compared to human-written reviews (Yao et al. 2017). On the other hand, the cost of machine-generated reviews is considerably lower than buying from sellers of human-crafted fake reviews. One study reported an average commission of $6.24 for buying human-written fake reviews (He et al. 2022). Whereas, GPT models such as ChatGPT, which are experiencing mass adoption (Heikkilä 2022), may be prompted to maliciously generate fake reviews at no cost or at a minimal fraction of that cost (see prices in the *Fake Reviews Generation* Section). In addition, GPT models present a high level of accessibility and user-friendliness, as they can simply be accessed by creating an account at OpenAI.com, using its playground at https://platform.openai.com/playground.

Overall, there is a large body of literature aimed at detecting, explaining, and analyzing how user-generated fake reviews impact both customers and social media platforms (Paul and Nikolaev 2021; Wu et al. 2020). However, no effort has been directed toward explaining, under different heterogeneous metrics, how machine-generated reviews are present on such platforms thus far, leaving a critical research gap in the literature. This paper aims to close such a research gap. We conjecture that machine-generated reviews may have a fraudulent nature because (1) they lack the authenticity and personal touch of a genuine review, *i.e.*, the content of the review might not accurately represent the users' experience, thus not providing accurate feedback, and (2) they have a limited perspective, lacking the contextual knowledge and experience that humans have, meaning that machine-generated reviews may not take into account factors that are important to the users such as their preferences, past experiences, and expectations. For these reasons, machine-generated reviews can be considered a subset of fake reviews aimed at distorting customer experiences. Formally, we define *AI-generated* fake reviews as reviews that are generated by AI systems using deep learning and natural language processing (NLP) techniques rather than by actual customers. For clarity, we will interchangeably use the terms *AI-generated*, *machine-generated*, and *GPT-generated* to refer to this concept for the rest of the paper.

In summary, this paper aims to address the following empirical research questions:

RQ1: How can *AI-generated* fake reviews fabricated from high quality *user-generated authentic* reviews be effectively detected?

RQ2: How do *AI-generated* fake reviews and *user-generated authentic* reviews differ across multiple dimensions of review, user, restaurant, and content characteristics?

In this study, we leverage Yelp's verified elite reviews as a basis for generating synthetic fake reviews using OpenAI's GPT-3 model. We conduct an online survey to test the human ability to distinguish between AI-generated and user-generated authentic reviews. The results reveal that participants struggle to differentiate between the two types of reviews, likely due to cognitive limitations in detecting patterns from large-scale data (Kahneman and Tversky 1972). In contrast, we train multiple fake review detection algorithms and discover that a GPT-3 model fine-tuned on our proposed dataset achieves the best performance, with an F1-score of 95.48%. Furthermore, instead of relying on filtered reviews as proxies for fake reviews commonly used in existing literature (Luca 2016; Luca and Zervas 2016; Rayana and Akoglu 2015), we opt for a more rigorous evaluation of non-filtered reviews, *i.e.*, reviews that pass the Yelp's filtering system that considers them as authentic, and are shown on the business pages. Our analysis focuses on potential fake reviews that

have evaded detection by Yelp's filtering system. To comprehensively assess these reviews, we examine them across various dimensions, encompassing review characteristics, user characteristics, restaurant attributes, and writing style.

This paper has the following contributions. Firstly, in line with previous literature on user-generated fake reviews, we find that machine-generated reviews tend to exhibit a polarization towards higher star ratings (*e.g.*, 4-5 stars), and are predominantly posted by users with less established reputations, as evidenced by their limited review history. Secondly, AI-generated reviews are generally more comprehensible and exhibit less linguistic complexity regarding vocabulary usage than authentic reviews. Finally, a key contribution of our study is using foot-traffic mobility data to establish a correlation between AI-generated reviews and restaurant demand. Our findings reveal that machine-generated reviews are more frequently associated with restaurants that attract fewer customer visits, a reliable proxy indicator of overall demand. This connection between AI-generated fake reviews and customer visits has not been previously explored or established in the existing literature, making our findings valuable to the ongoing research on fake reviews and their impact on businesses.

Literature Review

Impact of Fake Reviews in Online Markets

Various economic agents, including retailers and platforms, are known to manipulate online reviews (Gössling et al. 2018; Lee et al. 2018). Driven by financial incentives, online merchants often distribute fake positive reviews for their own products or fake negative reviews against competitors' products (Crawford et al. 2015; Paul and Nikolaev 2021). For instance, increased review circulation positively impacted revenues for high-quality restaurants and negatively affected low-quality restaurants (Fang 2022). Moreover, online platforms tend to circulate fake reviews to boost website traffic, promote customer engagement (Lee et al. 2018), and increase revenue by generating sales (He et al. 2022). In some cases, individual users might also post fake content for reward-seeking purposes (Anderson and Simester 2014). Overall, fake reviews undermine informativeness and information quality (Zhang et al. 2017), diminish review credibility and helpfulness (Agnihotri and Bhattacharya 2016; Zhang et al. 2017; Zhao et al. 2013), and negatively influence new consumers' decision-making processes. Existing research has also shown that the proliferation of fake reviews increases consumer uncertainty (Zhao et al. 2013), leading to customer distrust towards online reviews (DeAndrea et al. 2018; Filieri et al. 2015; Zhuang et al. 2018) and reducing consumers' purchase intentions (Munzel 2016; Xu et al. 2020).

Fake Reviews Detection

Human evaluators have consistently struggled to differentiate user-generated fake reviews from genuine ones (Crawford et al. 2015). For instance, Ott et al. (2013) surveyed members of the general public to detect fake reviews, discovering that the best human judge achieved an accuracy of only 65%. Similarly, Ott et al. (2011) reported a 61% accuracy, while Plotkina et al. (2020) and Sun et al. (2013) recorded human performance at average accuracy detection rates of 57% and 52%, respectively. These results indicate that human evaluators perform at an accuracy level comparable to random guessing. In contrast, automated detection of user-generated fake reviews, framed as a binary "*spam* versus *non-spam*" or "*fake* versus *non-fake*" supervised learning problem, has shown promising results (Paul and Nikolaev 2021). Benchmark models such as logistic regression (Liu et al. 2019) and random forest (Zhang et al. 2016) have served as a foundation for more advanced models like deep convolutional neural networks and recurrent neural networks (Zhang et al. 2018). Furthermore, LLMs such as RoBERTa have been successfully employed in fake review detection tasks, achieving F1 scores as high as 97% (Salminen et al. 2022).

Fake Reviews Characteristics

Writing Style

Korfiatis et al. (2008) posited that reviews' readability serves as a proxy for their helpfulness, as consumers must first read and then comprehend the text to assess their usefulness. Empirical research has further demonstrated that the likelihood of a review being deemed helpful increases when it is presented in an easily comprehensible manner (Cao et al. 2011). Hence, several studies theorized that fraudsters might deliberately disseminate simple fake content to quickly catch readers' attention (Agnihotri and Bhattacharya 2016; Li et al. 2013), conceptualizing that fake reviews were easier to comprehend. Empirically, leveraging readability metrics such as the *Automated Readability Index* (Senter and Smith 1967), Harris (2012) found that fake deceptive reviews exhibited less writing complexity as compared to truthful ones. However, no unanimous academic consensus has been established on this finding, because other studies employing comparable methodologies showed the opposite result (Banerjee and Chua 2014; Yoo and Gretzel 2009). Also, textual review sentiment has been investigated for its effectiveness and helpfulness (Tang et al. 2014). As for user-generated fake reviews, consumers realized that more polarized sentiment tones could be surrogates for suspicious user-generated content (Liljander et al. 2015). For example, prior research discovered that fake reviews were richer in positive cues as compared to authentic ones (Banerjee and Chua 2014; Yoo and Gretzel 2009). Besides, spammers were discovered not capable of expressing true sentiment when writing fake reviews, eventually leading to more polarized opinions (Liu and Pang 2018).

Ratings and Restaurant Characteristics

Extreme sentiment polarity was also detected when considering review ratings (Luca and Zervas 2016), which are robust complements of review textual sentiment (Cho et al. 2022). In particular, extant literature affirmed that positive fake reviews were more prevalent than negative ones (Lappas et al. 2016; Zhang 2019). For example, Lappas et al. (2016) found that 56% of fake reviews were positive (4-5 stars) and that 29% were negative (1-2 stars). One hypothesis that may ex-post explain the prevalence of positive fake content could be that a one-star increase in the Yelp restaurant average rating is associated with a 5-9% revenue growth (Luca 2016). As for restaurant characteristics, Luca and Zervas (2016) examined how fake restaurant reviews were present on Yelp. They found that about 16% of the reviews were filtered out as fake or suspicious, and that restaurants with fewer associated reviews were more likely to submit positive fake reviews to enhance their reputation. Luca and Zervas (2016) also segmented restaurants into chain (*e.g.*, McDonald's, Burger King, Subway, etc.) and non-chain, finding the former ones less likely to display positive fake content, because their revenue is not significantly affected by their rating (Luca 2016), and because they may incur high reputation costs if caught (Mayzlin et al. 2014).

User Characteristics

Fake reviews can also be identified by user behavior, *i.e.*, spammers' characteristics. For example, Sandulescu and Ester (2015) describe the concept of *singleton reviews*, which is the phenomenon of users posting only one (fake) review. Because of that one-to-one relationship, spotting and tracking activities of singleton review spammers is challenging (Rayana and Akoglu 2015). Barbado et al. (2019) defined four subsets of user-centric features to analyze Yelp reviews: *personal profile* features (*e.g.*, profiles description), *social interaction* features (*e.g.*, user number of friends), *review activity* features (*e.g.*, number of previous reviews), and *trust information* features (*e.g.*, number of photos posted). Here, they leveraged supervised machine learning techniques to classify fake versus authentic reviews, showing that *review activity* features were the most relevant in terms of classification accuracy. Inherent to our paper, they also described how accounts associated with consistent spamming of fake user-generated content displayed fewer friends, fewer photos posted, and fewer reviews as compared to accounts conducting a genuine activity. Similarly, Luca and Zervas (2016) found congruent results.

Research Design

This section describes the methodology employed to answer the study's RQs. We outline how: (1) collected data and generated fake GPT-3 reviews, (2) asked human judges and implemented machine learning algorithms to detect them, and (3) inferred and explained the predictions of machine-generated and authentic reviews on a set of unverified reviews. Figure 1 illustrates the GPT-3 pipeline from fake review generation to detection. As of 2022, GPT-3 is a state-of-the-art LLM developed by OpenAI that has gained considerable attention due to its impressive performance in a wide range of language-related tasks (Brown et al. 2020). Contextually, its applications such as ChatGPT are experiencing mass adoption (Heikkilä 2022). GPT-3 learns language patterns on an unprecedented scale, outputting human-like text. Generating text using LLMs is referred to as "prompt engineering", which involves optimizing the input called the *prompt* for the desired model response. It involves selecting length, language, and context for relevant, accurate, and useful output (Liu et al. 2023). Effective prompt engineering is crucial for high-quality content such as machine-generated fake reviews (Ouyang et al. 2022). In summary, the power and versatility of GPT-3 make it a valuable methodology for both text generation and detection.

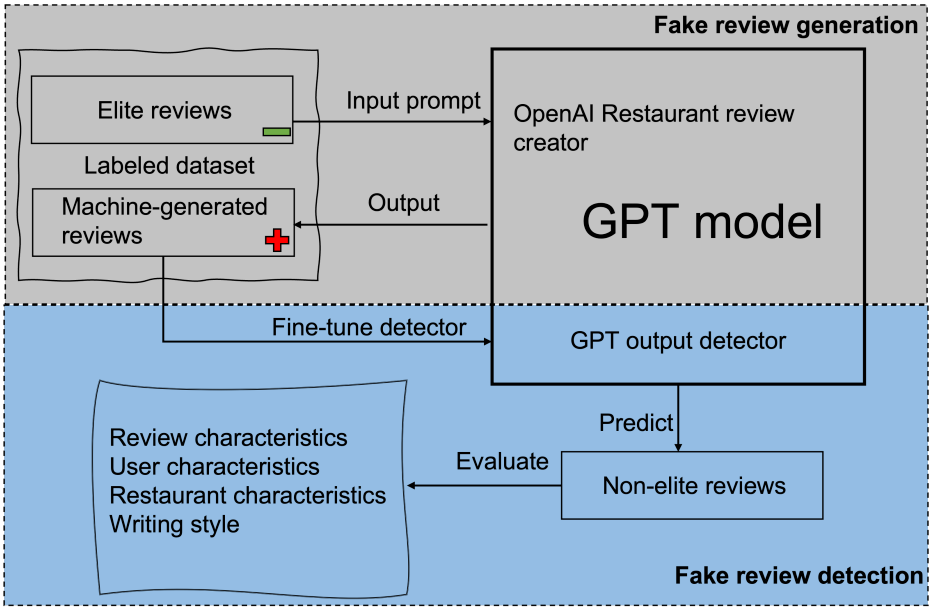

Figure 1. GPT-3 Methodology From Fake Review Generation To Detection

Data Collection

We accessed the 2021 to mid-2022 New York City restaurant mobility data from the company SafeGraph to collect a dataset of restaurants (https://www.safegraph.com). New York City was selected because it offers a variety of restaurants serving distinct culinary tastes within an international setting, thereby providing sufficient heterogeneity for our study. Next, we scraped all restaurant-related customer reviews from Yelp that had already passed its filtering system after 2020. To mitigate misinformation, Yelp has been implementing a filtering system to display the most reliable reviews on restaurant web pages (McCluskey 2022), making it a trustworthy preliminary gate to block fake content (Mukherjee et al. 2013). However, OpenAI released its API in late 2020, so we conjectured that "AI crowdturfing" campaigns were implemented after that year, given the API's easy accessibility and low usage costs (price rates in the next section). "AI crowdturfing" may be defined as campaigns that use AI systems (such as automated bots) to produce and distribute fake reviews with the aim of manipulating online reputations (Tricomi et al. 2022; Yao et al. 2017). In total, we

collected 177,410 reviews connected to 5,959 restaurants, divided into 131,266 non-elite and 46,144 elite reviews. Each example includes the review text, the date it was posted, the rating, the poster's Yelp elite status, the poster's number of previous reviews, and the poster's number of uploaded photos. Then, we enriched the data by (1) querying the Yelp official API downloading restaurant-related variables (each review was connected to), such as the average rating and the price level, and (2) including the raw number of visits and the normalized number of visits by the total visits (in the New York state) from the original SafeGraph data. SafeGraph collects visit data by leveraging various data sources, including GPS, Wi-Fi, and Bluetooth signals from visitors' mobile devices. The company uses a combination of these signals to determine the temporal location of devices in the physical world. Then it aggregates this information to create a comprehensive dataset of visit information. Overall, SafeGraph datasets have been extensively used in diverse research domains, including public health (Chang et al. 2022), and impact of mobility restrictions and compliance (Charoenwong et al. 2020), among others.

Fake Reviews Generation

After collecting the main data, we used the OpenAI publicly available GPT-3 API to build a dataset of fake reviews (https://openai.com/api). As of 2022, four different GPT-3 sub-models could be chosen at different price rates: *Ada* (0.0004$ / 1K tokens), *Babbage* (0.0005$ / 1K tokens), *Curie* (0.0020$ / 1K tokens), and *Davinci* (0.0200$ / 1K tokens). A token roughly corresponds to one English word. Naturally, the higher the price rate, the more accurate the model instruction-following. A higher price rate also translates into larger model sizes, as measured by the number of parameters, which are not officially disclosed by OpenAI. Simply put, a larger number of parameters leads to increased performance. We then randomly sampled 12,000 reviews from the 46,144 elite reviews representing 4,994 restaurants and used the elite-sampled texts as prompts to generate related fake reviews. Incorporating elite reviews into the prompt aims to generate machine-generated fake reviews that closely mimic their sophistication, *i.e.*, high-quality, as our analysis focuses on reviews that have evaded detection from the Yelp filtering system. Elite reviews are written by elite users, who Yelp thoroughly verifies (Zhang et al. 2020), and their reviews are accompanied by an elite badge, which indicates the reviewer's status at the time when the data was collected. Users are incentivized to obtain an elite status to access new features on the Yelp platform, as well as special offers, discounts, or promotions from local businesses, among others (Wang et al. 2021). According to Yelp, to apply for an elite membership, a user is expected to have consistently posted thoughtful reviews, uploaded beautiful pictures, and up-voted others' reviews. Therefore, we assume that elite reviews are a reliable proxy of information reflecting real customers' opinions. Practically, to fabricate synthetic fake reviews, we utilized the default prompt that OpenAI provides for restaurant fake review generation (https://platform.openai.com/examples/default-restaurant-review):

> *"Write a restaurant review based on these notes:*
> *Name: <EXAMPLE RESTAURANT NAME>*
> *<EXAMPLE ELITE REVIEW TEXT>"*

The adoption of the default prompt allows for a favorable compromise between easy accessibility and high text generation quality. For each fake review generation, we randomly selected one model between *Curie* and *Davinci* with equal probability. Also, we randomly sampled the *Temperature* value, a hyper-parameter controlling the randomness of the generated text, from $U \sim (0.3, 0.7)$. A value of 0 generates a deterministic and repetitive text and 1 vice versa. All the other hyper-parameters were kept as default. For later use, the final dataset of 24,000 reviews, equally balanced between authentic and fake reviews, was split into 80% training and 20% testing.

Fake Reviews Detection

We employed (1) members of the general public and (2) implemented machine learning solutions to detect machine-generated fake reviews.

Firstly, we ran a survey in which each respondent was asked to select the machine-generated review from a set of review pairs. We sampled 15 review pairs from the training set to "train" the respondents for the task, with each pair containing one human-generated review and one AI-generated review and related ground-

truth answers. Similarly, we sampled 40 review pairs from the test set and used those as survey questions. Participants were shown the training examples before accessing the questions. Additionally, the reviews were paired assuring length comparability (max 30 words difference) and sentiment comparability at the aggregate level, meaning that the same number of positive, neutral, and negative reviews occurred both for fake and real reviews. For completeness, reviews in the dataset are about 140 words in length on average (std 75). In relation to the response options, a third alternative, *"Cannot decide. I'm unsure"*, was included to enable study participants to indicate their uncertainty instead of having to resort to a random guess. This additional response choice aimed to enhance the accuracy and reliability of the survey data collected by reducing the impact of arbitrary guessing, which may compromise the validity of the study. The order of the questions was randomly spread across the survey form, and two questions were converted into attention checks to monitor the respondents' care. In the attention checks, we explicitly solicited participants to select one option. The survey was then sent to 90 random participants in the US through the Prolific platform, and they were paid about $8.30 per hour. After removing 10 attempts because of at least one inattentive answer, we counted 80 valid responses related to 38 questions, totaling 3,040 single-question responses.

Secondly, as for machine learning solutions, we fine-tuned a pre-trained GPTNeo model to classify fake versus real reviews. GPT models belong to the family of transformer models, which have become state-of-the-art in NLP and computer vision. The reason why transformer models are powerful is that they rely on the attention mechanism (Vaswani et al. 2017), allowing the network to mainly focus on the most relevant parts of the input sequence. GPTNeo is designed using EleutherAI's replication of the GPT-3 architecture, which currently is OpenAI's proprietary software. As such, GPTNeo is a scale-up of the GPT and GPT-2 models (Radford et al. 2019). Practically, we accessed a 125 million parameters pre-trained version from Huggingface (https://huggingface.co/EleutherAI/gpt-neo-125M), and fine-tuned it with our generated fake restaurant reviews dataset. We benchmarked GPTNeo with other machine learning models such as Bidirectional-LSTM (BiLSTM), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), XGBoost (XGB), and GPT-2. We also benchmarked it to the current official open-source OpenAI's RoBERTa model for GPT fake text detection (Solaiman et al. 2019). For LR, NB, RF and XGB we trained with 5-fold cross-validation on the training set and reported the results on the test set. Here, review texts were represented with a bag-of-words approach, which tokenizes text into individual words and then counts the frequency of those words in each document. While for the deep learning models (BiLSTM, GPT-2, and GPTNeo), we extracted another 20% partition from the training set as validation data since computing 5-fold cross-validation is computationally expensive. Here, review texts were tokenized using Byte-Pair Encoding (BPE), which is a byte-level data compression algorithm used to segment words into subword units by iteratively merging the most frequently occurring pairs of adjacent bytes. We trained using the AdamW optimizer default hyper-parameters, using a learning rate of 1e-4, decaying it by a factor of 0.1 every 5 epochs, a batch size of 1, and early-stopping at 10 epochs. For the GPTNeo, we computed the optimal classification threshold at each epoch by optimizing Youden's J statistics in the validation set, calculated as the difference between the true positives rate and false positives rate (Salminen et al. 2022). Finally, the best weights and classification threshold were saved, and evaluation was performed on the test set. For all the models, we reported the accuracy score, precision score, recall score, and F1 score.

Inference on Non-Elite Reviews

With the best GPTNeo classifier, we performed inference on the 131,266 unverified non-elite reviews, determining the probability of each one being machine-generated. Each example review incorporates a *review*-based variable, *i.e.*, the review rating given by the review poster together with the review text (*Rating*), distributed as a 1 to 5 Likert scale; *user*-based variables, *i.e.*, the user's number of friends (*#Friends*), the user's number of previously posted reviews (*#Reviews*), and the user's number of previously posted photos (*#Photos*); and *restaurant*-based variables, *i.e.*, the restaurant's average rating computed by Yelp from all the reviews that passed its filtering system (*AvgRating*), the price level (*PriceLevel*), *i.e.*, the average price per person denoted as "$": under $10, "$$": $10–$30, "$$$": $31–$60 and "$$$$": over $60, the total number of reviews posted by customers (*#RestReviews*), the chain status (*ChainStatus*), computed adopting Zhang and Luo (2023) approach, which counts the number of unique restaurant names in the dataset, and assigns those appearing more than five times as belonging to a restaurant chain (*e.g.*, McDonald's, Starbucks, Burger King, etc.), the number of customer visits between 2021 and mid-2022 (*#Visits*), and the normal-

ized number of visits (*NormVisits*), multiplied by 1,000 for easier readability. Afterward, classification was performed with a sensitivity analysis approach at the [.5, .6, .7, .8, .9, .99, and J^*] classification thresholds. For each threshold t, we separated predicted machine-generated versus authentic reviews, and performed ANOVA for each aforementioned variable to inspect differences across the two predicted categories. This methodology was adopted because the labels about whether non-elite reviews were AI-generated were not available. Thus, different thresholds were tested to examine the sensitivity and robustness of predictions.

Finally, we conducted a robustness test to evaluate GPTNeo's accuracy in classifying generated reviews from older GPT models. We further collected 12,000 elite pre-GPT-3 reviews representing the same restaurants and used a GPT-2 model to generate an equivalent number of fake reviews based on the prompt illustrated in Subsection *Fake Reviews Generation*. Then, GPTNeo was employed for the classification of both the elite (authentic) and GPT-2-generated (fake) reviews. This test provides increased confidence in GPTNeo's ability to identify reviews produced by older GPT models, thus making the analysis more temporally robust.

Writing Style: Explaining the Predictions

In our context, writing style refers to how a textual review is constructed by the writer, sentence-by-sentence, and word-by-word. We believe that humans and machines have different writing styles, with the latter being more repetitive, more predictable, and less sophisticated than the former. We considered three classes of metrics to evaluate the writing style of each non-elite review: *perplexity*-based, *readability*-based, and *sentiment*-based metrics. *Perplexity*-based metrics include *Perplexity* (*PPL*) and *Textual Coherence* (*TC*). As in Equation 1, *PPL* is defined as the exponential average negative log-likelihood of a sequence of words $w_i, w_{i+1} \dots w_{i+t}$.

$$PPL(W) = \exp\left[-\frac{1}{t}\sum_{1}^{t}\log p(w_i|w_{<i})\right] \tag{1}$$

In simple terms, it measures the conditional probability that each word follows its preceding one. *PPL* is one of the most widely adopted metrics to optimize and evaluate the accuracy of LLMs, with low *PPL* values implying better accuracy. Notably, LLMs tend to generate common words more often as opposed to unusual vocabulary (Heikkilä 2022). Therefore, unless carefully prompted, LLMs may generate text that is less sophisticated than human-written one in terms of lexical terminology, outputting text with relatively low *PPL* values. Next, by breaking a review into a sequence of sentences, we introduce *Textual Coherence* (*TC*). *TC* is defined as the presence of semantic relations among sentences. In simple words, given a corpus containing a set of sentences that when viewed independently convey a valid meaning, if by reading them sequentially no meaning is conveyed, then the corpus is not coherent. To measure *TC*, we deployed the *Zero-Shot Shuffle Test* (Laban et al. 2021). For each review, we generated many random sentence permutations, scored each of them using Equation 1, and subtracted the original *PPL* score, obtaining a per-review set of perplexity changes, which were averaged to determine *TC*. Some reviews were lengthy, causing a significant computational cost due to permutation generation. To cope, we selected a subset of s sentences such that $s = min(n, 5)$. This was done to avoid the high computational cost of generating all possible permutations ($O(n!)$), and because 5 per-review permutations are still computationally feasible to be scored. Finally, *PPL* and *TC* were calculated using a general purpose pre-trained 125-million-parameter GPTNeo model.

As for *readability*-based metrics instead, we considered the following metrics: *Automated Readability Index* (*ARI*) (Senter and Smith 1967) and *Number of Difficult Words* (*#DW*). *ARI* is one of the most widely adopted readability indices to evaluate the readability of a given text. In particular, it has already been used to evaluate the readability of online reviews (Harris 2012; Hu et al. 2012). As in Equation 2, *ARI* decomposes the text into basic structural elements such as the number of characters (*#Chars*), number of words (*#Words*), and number of sentences (*#Sentences*).

$$ARI = 4.71\frac{\#Chars}{\#Words} + 0.5\frac{\#Words}{\#Sentences} - 21.43 \tag{2}$$

Unlike other readability indices, the main advantage of *ARI* is that it relies on the number of characters per

word and not on the number of syllables per word, therefore being more accurate to calculate for a computer. Also, the interpretation of *ARI* is straightforward, as its output produces an approximated representation of the US-grade education level needed to understand the text. For example, an *ARI* of 9.2 indicates that a 9th-grade student can understand the text. Simply put, the higher the *ARI* score, the higher the difficulty in text comprehension for an average interlocutor. Next, *#DW* is the count of difficult words present in a text. By looking at the *Dale-Chall Word List* (Dale and Chall 1948), which contains approximately 3,000 familiar words known by an average 5th-grade student, if a word is not present in the list, then it is labeled as difficult. Finally, the only *sentiment*-based metric considered is the SiEBERT *Sentiment* score (Hartmann et al. 2023). SiEBERT is based on a RoBERTa architecture and fine-tuned on 15 different datasets. Its output ranges from -1 (negative) to +1 (positive).

To sum up, we scored each review with the *perplexity*-based, *readability*-based, and *sentiment*-based metrics. Afterward, for each metric, we performed ANOVA to inspect differences across the predicted machine-generated versus real reviews at each classification threshold t as earlier introduced.

Results

Human Evaluations versus Model Evaluations

Surveyed people from the general public only attained an average accuracy score of 57.13% (std 13.57%) net of the answers they refrained from answering, meaning that humans were only 7.13% better than random guessing (=50%) in our experimental setting. Besides, we recorded an average abstention rate (*i.e.*, selecting the *"Cannot Decide. I'm unsure"* option) of 11.15% (std 12.66%). Therefore, humans were not effectively capable of distinguishing GPT-3 machine-generated content from user-generated one. Conversely, machine learning algorithms attained significantly better performance than human evaluators. In Table 1 we provide the classification report of the classifiers.

Model	Accuracy %	Precision %	Recall %	F1-score %
OpenAI	76.78	84.87	64.98	73.60
NB	83.48	93.62	71.72	81.22
RF	82.39	84.61	79.01	81.72
XGB	83.18	86.59	78.38	82.28
LR	85.07	87.00	82.34	84.61
BiLSTM	93.71	93.09	93.09	93.09
GPT-2	94.63	94.57	94.65	94.61
GPTNeo	95.21	94.74	**95.70**	95.22
GPTNeo@J	**95.51**	**95.80**	95.15	**95.48**

Table 1. Classification Report On The Test Set

Surpassing all the benchmarks, GPTNeo models ranked as top performers. Specifically, the GPTNeo maximizing accuracy after calculating the Youden's J statistics as the optimal classification threshold in the validation set achieves the best performance (GPTNeo@J). Convergence occurred at the 2nd epoch, with optimal J^*=.5708. Overall, GPTNeo@J significantly outperforms human evaluators and OpenAI's official benchmark by 38.38% and 18.73% accuracy, respectively. Finally, coherent with the accuracy in classifying GPT-3-generated fake reviews, we recorded 94.41% and 96.59% GPTNeo@J accuracies in classifying GPT-2-generated fake reviews and their respective elite counterparts used for generation, respectively. With these findings, we applied the optimized GPTNeo@J model for inference on the unverified non-elite reviews.

ANOVA Results

Unless differently specified, ANOVA results are discussed at the significance level α = .05 and at the optimized classification threshold J^* =.5708. In Table 2, we provide a per-variable summary with averages for predicted human-written reviews and AI-generated fake reviews, respectively. Out of a total of 131,266 non-elite reviews posted from 2021 onward, 8.48% were predicted as machine-generated. This percentage monotonically decreases as the threshold t is increased. For instance, at t=.99, only .10% of reviews were

detected as AI-generated.

Name	Category	Humans	AI	F-statistic
Rating	Review	3.94	4.37	914.91***
#Friends	User	71.19	65.60	10.45**
#Reviews	User	44.21	32.84	78.46***
#Photos	User	66.81	32.49	22.96***
AvgRating	Restaurant	3.97	4.00	41.84***
PriceLevel	Restaurant	2.24	2.22	3.52
#RestReviews	Restaurant	743.68	788.39	18.61***
#Visits	Restaurant	3004	2866	5.22*
NormVisits	Restaurant	0.19	0.18	5.18*
ChainStatus	Restaurant	0.14	0.14	0.04
Perplexity	Writing	78.70	83.38	14.01***
Coherence	Writing	25.31	21.20	3.33
ARI	Writing	7.05	6.82	13.35***
#DW	Writing	10.70	6.76	1847.82***
Sentiment	Writing	0.47	0.71	779.17***

Table 2. ANOVA (J^*=.5708). *p<.05, **p<.01, ***p<.001

Review-based and User-based

For the review *Rating*, and the users' *#Reviews*, *#Friends* and *#Photos* all differences were statistically significant. Here, reviews classified as machine-generated were given a higher average star *Rating* (+.43, p<.001). As for user-based variables, predicted machine-generated reviews were posted by users with a lower average number of *#Friends* (-5.59, p<.01), a lower average number of previously posted *#Reviews* (-11.37, p<.001), and a lower average number of previously posted *#Photos* (-34.32, p<.001). In Figure 2, we show a sensitivity analysis considering other thresholds t of classification (yellow and green subplots). Overall, *Rating* and *#Reviews* exhibited statistically significant up-trend and downtrend divergences for each t, respectively, while *#Photos* and *#Friends* no longer demonstrated statistical significance at t=.99 and from t=.6, respectively.

Restaurant-based

We observed that predicted machine-generated reviews were associated with restaurants with a higher *AvgRating* (+.03, p<.001). However, we acknowledge the modest practical implications that such a minimal difference may bring about. Then, we documented statistical significance for *#RestReviews*. Predicted fake reviews were connected to restaurants that displayed a greater average number of reviews available (+44.71, p<.001). The opposite was observed for the average *#Visits*, in which predicted AI-generated reviews were linked to restaurants that received fewer customer visits from 2021 to mid-2022 (-138, p<.05). This result was strengthened by *NormVisits* (-.01, p<.05). Finally, no significant differences were noticed for the *ChainStatus* and the *PriceLevel* (p>.05). In Figure 2, we show a sensitivity analysis considering other thresholds t of classification (blue subplots). Overall, *#RestReviews*, *#Visits* and *NormVisits* exhibited statistical significance for each t, except for t=.99.

Writing Style

AI-generated fake reviews showed a higher average *Perplexity* (+4.68, p<.001). However, *Perplexity* also displayed a statistically significant downtrend when increasing the classification threshold t (p<.05), eventually scoring lower than for predicted human-generated reviews from t>.7. Instead, when considering *Textual Coherence* no statistical significance was observed (p>.05). Figure 2 shows the sensitivity analysis (orange subplots). As for *readability*-based metrics, predicted AI-generated fake reviews were discovered to be more readable and less difficult to comprehend compared to the human-generated ones. Both average *ARI* and average *#DW* scored lower for machine-generated content, (-.23, p<.001) and (-3.94, p<.001), respectively.

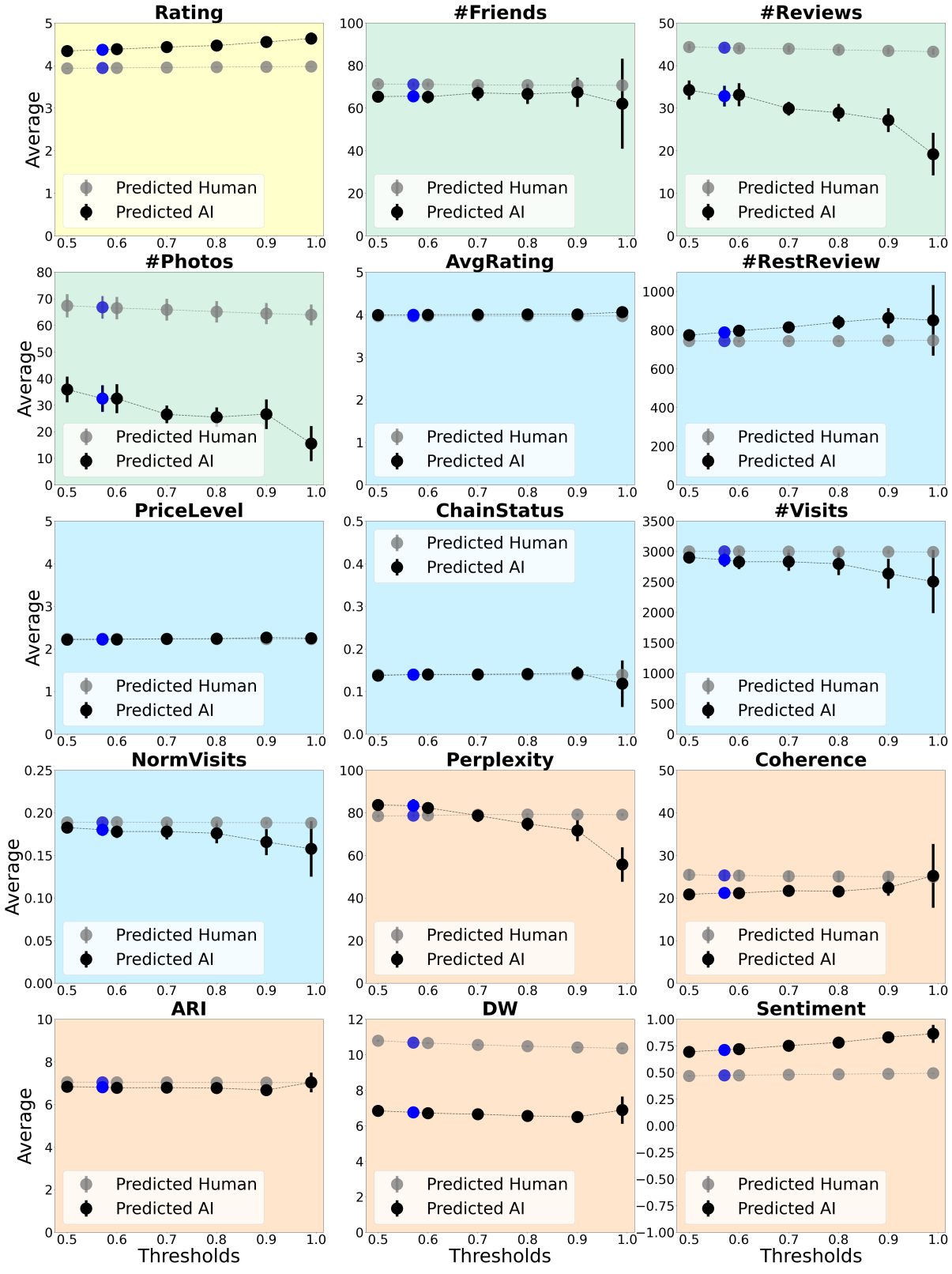

Figure 2. Sensitivity Analysis. Optimal J^*=.5708 Highlighted In Blue

Lastly, for *Sentiment*, we observed that predicted machine-generated reviews had a more positive tone (+.24, p<.001). Overall, *Sentiment* displayed a statistically significant up-trend divergence for each t. *ARI* was no longer statistically significant at t=.99. Finally, differences in *#DW* were statistically significant for each t.

Discussion

To address **RQ1**, we described how human evaluators systematically fail at detecting GPT-3 AI-generated content in the domain of restaurant reviews (57.13% accuracy), extending findings from studies that employed human judges to unsatisfactorily detect user-generated fake reviews to the realm of GPT-generated fake reviews ($\leq$ 65% accuracy rates) (Ott et al. 2011; Plotkina et al. 2020; Sun et al. 2013). In other words, humans are not capable of distinguishing neither user-generated nor machine-generated fake reviews. In contrast, AI-generated fake reviews can be effectively detected using models based on the equivalent GPT structure (GPTNeo@J, +38.38% accuracy, 4.49% out-of-sample error on our crafted dataset). Such disparity in performance between humans and machines could possibly be attributed to human cognitive limitations at detecting patterns from large-scale unstructured data (Kahneman and Tversky 1972). In our experiment, before presenting the survey questions, we "trained" the respondents by showing 15 question-answers pairs from the same distribution to give context for the task. However, human learners encountered challenges when attempting to acquire new knowledge within the complex and unfamiliar context of fake reviews detection. Specifically, one major limitation may be attributed to the limited attention span of humans, which, in turn, may lead to information overload (Navon and Miller 2002). Additionally, confirmation bias can influence pattern recognition by causing individuals to overlook or dismiss patterns that do not align with their earlier established (wrong) expectations (Nickerson 1998). Yet, another challenge may be posed by the presence of complexity and noise in the semantic structure of online reviews (Wickens et al. 2015). To cope, deep learning algorithms helped to address complexity of textual data by automating the process of pattern recognition, which is beneficial when dealing with large volumes of data (*e.g.*, in fake review detection tasks).

We then applied the best GPTNeo@J model (J^*=.5708) to our sample of unverified customer reviews published after 2020 that had already passed the Yelp filtering system, documenting that 8.48% of them were predicted as machine-generated. In comparison, prior research on Yelp reported filtering out user-generated fake reviews at estimated rates around 16% (Luca and Zervas 2016). However, it would be misleading to conclude that the incidence of machine-generated fake reviews is about half that of user-generated fake reviews, because (1) our GPTNeo@J has an error rate of 4.49% on the test set, and (2) Yelp does not publicly open-source the algorithm behind its filtering system. Therefore, the impossibility of directly comparing the two filtering methodologies, added to the GPTNeo@J error rate, is one limitation of our study. Lastly, we conducted a robustness test by applying GPTNeo@J on a set of 12,000 GPT-2-generated fake reviews and the elite counterparts used for their generation, recording 94.41% and 96.59% accuracies, respectively. These results align with GPTNeo@J accuracy on the test set for GPT-3-generated reviews detection (95.51%), demonstrating that GPTNeo@J effectively identifies fake reviews generated from older GPT models, thus making the analysis more temporally robust.

Firstly, we observed that machine-generated reviews score a higher average *Rating* compared to the human-generated ones (+.43, p<.001), with an upward divergence at each sequential t strengthening our result. This finding is congruent with conclusions on user-generated fake reviews from Luca and Zervas (2016), who documented that user-generated fake reviews have a bimodal distribution with spikes at 1 and 5 stars, and from Lappas et al. (2016), who singled out 56% of positive reviews (4-5 stars) out of 15,000 Yelp fake reviews. Extant literature may justify our result, suggesting that economic agents seeking to bolster or restore their reputation may be more likely to engage in the self-promotion of falsely positive reviews (Luca and Zervas 2016) because a 1-star increase in the Yelp average rating is linked with a 5-9% revenue growth (Luca 2016).

Secondly, consistent with prior observations on user-generated fake reviews (Barbado et al. 2019; Luca and Zervas 2016), users that post machine-generated reviews have less established Yelp reputations as compared to those that allegedly post real content: fewer previously posted *#Reviews* (-11.37, p<.001), fewer *#Friends* (-5.59, p<.01), and fewer previously posted *#Photos* (-34.32, p<.001). Such diminished engagement levels might indicate an inclination toward spamming activities. It might be logical to presume that fraudsters engaging in spamming behavior would demonstrate lower levels of activity on a given platform. This might

be due to their employment of rotating accounts to disseminate fabricated content, which might justify their lack of interest in cultivating reliable and trustworthy reputations within the Yelp community.

Thirdly, regarding *restaurant*-based variables, our study found no significant impact of AI-generated reviews on the overall average restaurant rating, price level, and chain status. Specifically, the very marginal difference of +.03 (p<.001) in *AvgRating* lacks practical relevance since humans are not affected by such a small difference. Moreover, no difference in AI-generated reviews from either *PriceLevel* or *ChainStatus* showed any statistical significance. However, our findings on machine-generated reviews are in contrast with prior research from user-generated fake reviews concerning chain restaurants, as Luca and Zervas (2016) found that they are less likely to display fake content to protect their brand reputation, and the number of total reviews (*#RestReviews*, +44.71, p<.001). Yet, Luca and Zervas (2016) posited that restaurants have a stronger incentive to post fake reviews when few reviews are available, because the marginal benefit of each additional review is higher as Yelp displays the average rating as an indicator of customer satisfaction.

Interestingly, by leveraging the power of the SafeGraph data, which reports the estimated per-restaurant number of customer visits (*#Visits*), we concluded that restaurants that displayed more machine-generated reviews totaled fewer customer visits (-138, p<.05). To the best of our knowledge, this study represents the first analysis leveraging real users' visits to describe how fake reviews correlate with customer visits in the hospitality sector. Also, this finding raises novel research questions to investigate the influence of fabricated reviews on business performance. This research direction is motivated by the need to gain a deeper understanding of the potential effects of fake reviews on consumer behavior, which can inform business strategies and policies to promote transparency in online marketplaces.

Finally, we inspected the writing style of the two predicted review categories. As for *perplexity*-based metrics, our results suggest that perplexity exhibits a downtrend pattern when applied to sequential thresholds of classification t. Specifically, our findings indicated that, at t<.7, the average *Perplexity* of predicted AI-generated reviews was higher than for human-generated text (p<.01), whereas it was lower at t>.7 (p<.05). Additionally, *Textual Coherence* was not found to be statistically significant at any threshold. The pattern of *Perplexity* may be explained by analyzing how LLMs are developed and generate text. LLMs are trained by predicting the next most likely token in a sequence of words, minimizing textual perplexity, being more likely to output common words instead of rare words (Heikkilä 2022). Thus, it is reasonable to assume that AI-generated texts have lower perplexity in comparison to human-generated ones, meaning that LLMs demonstrate reduced uncertainty in generating text. In other words, perplexity may reflect the likelihood of a text being machine-generated, with lower values indicating a higher probability of machine generation. Our study reports a statistically significant downtrend in *Perplexity* for AI-generated text across all thresholds t (see Figure 2). Higher values of t can be interpreted as a higher level of confidence in classifying text as AI-generated. Therefore, we conjecture that as our confidence in classification increases, the likelihood of misclassifying an AI-generated text decreases, thus leading to lower perplexity. Our findings are consistent with this conjecture. In practical terms, because of lower *Perplexity*, AI-generated reviews exhibit greater word predictability, yet they may lack word originality and creativity, as well as potentially be repetitive.

Next, as for *readability*-based metrics, we showed that machine-generated reviews bear a higher degree of comprehension, necessitating a lower educational grade to be understood, as measured by *ARI* (-.23, p<.001) and *#DW* (-3.94, p<.001). These findings are congruent with Harris (2012), but different from Yoo and Gretzel (2009) on user-generated fake reviews. In our sample, predicted human-generated reviews and AI-generated reviews score *ARI* values of 7.05 and 6.82, respectively, meaning that they can be understood by average 7th and 6th-grade US students, respectively. Written content that is easily comprehensible can reduce the cognitive load on readers' information processing capabilities, potentially attracting a larger readership and positively affecting the perceived helpfulness of reviews (Agnihotri and Bhattacharya 2016). As for sentiment, aligned with Banerjee and Chua (2014), Liu and Pang (2018), and Yoo and Gretzel (2009), AI-generated fake reviews presented a more positive tone (*Sentiment*, +.24, p<.001). Spammers adopting AI-solutions may be deliberately employed to alter customers' perceptions by prompting GPT models to use exaggerated language that translates into more polarized (positive) sentiment polarities. This is also congruent with our previous result that AI-generated fake reviews tend to have higher average ratings (+.43, p<.001), as both variables are closely intertwined (Cho et al. 2022).

This study is not without limitations. Firstly, we restricted the analysis to the city of New York to avoid sampling biases. However, we cannot conclude whether the results can be generalized to other areas. Secondly, we relied on the SafeGraph dataset to collect restaurant reviews. Yet, SafeGraph does not disclose the exact data collection methodology for selecting example restaurants, leaving us with uncertainty about the representativeness of the dataset. Thirdly, we only adopted one prompt template to generate fake reviews and used elite reviews as a source of information for GPT-3 during the generation phase. This strategy allowed us to fabricate high-quality machine-generated fake reviews, as demonstrated by the human inability to detect them even after being trained for the task. However, we acknowledge that spammers may have different expertise with prompting LLMs, thus the complexity of the generated content may accordingly vary. Fourthly, we only relied on 2021 and 2022 inferential data, because OpenAI's GPT API was proprietarily released in late 2020. However, years 2021 and early 2022 were still affected by the COVID-19 pandemic, thus weakening our results as compared to ordinary years, as local lockdowns may have been imposed by New York authorities, potentially changing customers' behavior. It may be reasonable to point out that results about the number of customers' visits may be subjected to changes during ordinary times. Fifthly, we highlight that when filtering out reviews at t=.99 only about 130 AI-generated fake reviews were singled out, reducing statistical power in ANOVA. Lastly and importantly, we do not provide any causal interpretation to the empirical results found both in **RQ1** and **RQ2**, preventing us from drawing *cause-effect* conclusions.

Conclusion

Disseminating fake reviews has become more accessible and cheaper than ever after advances in LLMs such as ChatGPT. Such accessibility may amplify the proliferation of "AI crowdturfing" campaigns aimed at distorting user experiences on social media. This study proves that GPT machine-generated fake reviews can easily deceive readers, because of purported humans' cognitive limitations in dissecting machine-generated fake content from the authentic one. As the technology used to generate fake reviews becomes more advanced, review platforms must keep pace with technological advancements to ensure they can detect and remove such content effectively. To this end, we implemented AI-based detection and description of machine-generated fake reviews that had already passed the Yelp filtering system across review-based, user-based, restaurant-based, and writing style characteristics, showing that fake reviews tend to have a higher rating, that users posting more AI-generated content have less established Yelp reputations, and that such AI-generated reviews are easier to understand and less sophisticated in terms of vocabulary as compared to the human-generated ones. These findings are mostly aligned with conclusions from extant literature on user-generated fake reviews, meaning that machine-generated fake reviews may serve the same purposes as fake human-generated ones do. Notably, without providing causal claims, we also described how restaurants displaying more machine-generated fake content are subjected to fewer customer visits. Up to now, no study has investigated how fake reviews correlate with customer visits as a proxy for restaurant demand. Thus, we also intend to open novel research questions in this direction.

Acknowledgements

This work was funded by Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020 and Social Sciences DataLab - PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016), and by Oracle. The authors warmly thank Rodrigo Belo for the feedback received.

References

Agnihotri, A. and Bhattacharya, S. 2016. "Online review helpfulness: Role of qualitative factors," *Psychology & Marketing* (33:11), pp. 1006–1017.

Anderson, E. T. and Simester, D. I. 2014. "Reviews without a purchase: Low ratings, loyal customers, and deception," *Journal of Marketing Research* (51:3), pp. 249–269.

Banerjee, S. and Chua, A. 2014. "A theoretical framework to identify authentic online reviews," *Online Information Review* (38:5), pp. 634–649.

Barbado, R., Araque, O., and Iglesias, C. A. 2019. "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management* (56:4), pp. 1234–1244.

Brown, T. et al. 2020. "Language Models are Few-Shot Learners," in: *Advances in Neural Information Processing Systems,* vol. 33. Curran Associates, Inc., pp. 1877–1901.

Cao, Q., Duan, W., and Gan, Q. 2011. "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," *Decision Support Systems* (50:2), pp. 511–521.

Chang, T., Hu, Y., Taylor, D., and Quigley, B. M. 2022. "The role of alcohol outlet visits derived from mobile phone location data in enhancing domestic violence prediction at the neighborhood level," *Health & Place* (73), p. 102736.

Charoenwong, B., Kwan, A., and Pursiainen, V. 2020. "Social connections with COVID-19–affected areas increase compliance with mobility restrictions," *Science Advances* (6:47), eabc3054.

Cho, H. S., Sosa, M. E., and Hasija, S. 2022. "Reading between the stars: understanding the effects of online customer reviews on product demand," *Manufacturing & Service Operations Management* (24:4), pp. 1977–1996.

Crawford, M., Khoshgoftaar, M. T., Prusa, D. J., Richter, N. A., and Al Najada, H. 2015. "Survey of review spam detection using machine learning techniques," *Journal of Big Data* (2:23), pp. 1–24.

Dale, E. and Chall, J. S. 1948. "A formula for predicting readability: Instructions," *Educational Research Bulletin* (27:2), pp. 37–54.

DeAndrea, D. C., Van Der Heide, B., Vendemia, M. A., and Vang, M. H. 2018. "How people evaluate online reviews," *Communication Research* (45:5), pp. 719–736.

Duan, W., Gu, B., and Whinston, A. B. 2008. "Do online reviews matter? — An empirical investigation of panel data," *Decision Support Systems* (45:4), pp. 1007–1016.

Fang, L. 2022. "The effects of online review platforms on restaurant revenue, consumer learning, and welfare," *Management Science* (68:11), pp. 8116–8143.

Filieri, R., Alguezaui, S., and McLeay, F. 2015. "Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth," *Tourism Management* (51), pp. 174–185.

Gössling, S., Hall, C. M., and Andersson, A.-C. 2018. "The manager's dilemma: a conceptualization of online review manipulation strategies," *Current Issues in Tourism* (21:5), pp. 484–503.

Harris, C. G. 2012. "Detecting deceptive opinion spam using human computation," in: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence,* AAAI'12. Toronto, Ontario, Canada.

Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. 2023. "More than a feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing* (40:1), pp. 75–87.

He, S., Hollenbeck, B., and Proserpio, D. 2022. "The market for fake reviews," *Marketing Science* (41:5), pp. 896–921.

Heikkilä, M. 2022. *How to spot AI-generated text*. MIT Technology Review. Accessed April 7, 2023.

Hu, N., Bose, I., Koh, N. S., and Liu, L. 2012. "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decision Support Systems* (52:3), pp. 674–684.

Hu, N., Liu, L., and Zhang, J. J. 2008. "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects," *Information Technology and Management* (9), pp. 201–214.

Jindal, N. and Liu, B. 2008. "Opinion Spam and Analysis," in: *Proceedings of the 2008 International Conference on Web Search and Data Mining,* WSDM '08. Palo Alto, California, USA: Association for Computing Machinery, pp. 219–230.

Kahneman, D. and Tversky, A. 1972. "Subjective probability: A judgment of representativeness," *Cognitive Psychology* (3:3), pp. 430–454.

Korfiatis, N., Rodriguez, D., and Sicilia, M. 2008. "The Impact of Readability on the Usefulness of Online Product Reviews: A Case Study on an Online Bookstore," in: *Emerging Technologies and Information Systems for the Knowledge Society: First World Summit on the Knowledge Society, WSKS 2008, Athens, Greece, September 24-26, 2008. Proceedings 1,* 2008, pp. 423–432.

Laban, P., Dai, L., Bandarkar, L., and Hearst, M. A. 2021. "Can Transformer Models Measure Coherence In Text: Re-Thinking the Shuffle Test," in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,* Online: Association for Computational Linguistics, 2021, pp. 1058–1064.

Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry," *Information Systems Research* (27:4), pp. 940–961.

Lee, S.-Y., Qiu, L., and Whinston, A. 2018. "Sentiment manipulation in online platforms: An analysis of movie tweets," *Production and Operations Management* (27:3), pp. 393–416.

Li, M., Huang, L., Tan, C.-H., and Wei, K.-K. 2013. "Helpfulness of online product reviews as seen by consumers: Source and content features," *International Journal of Electronic Commerce* (17:4), pp. 101–136.

Liljander, V., Gummerus, J., and Söderlund, M. 2015. "Young consumers' responses to suspected covert and overt blog marketing," *Internet Research* (25:4), pp. 610–632.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. 2023. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys* (55:9), pp. 1–35.

Liu, Y. and Pang, B. 2018. "A unified framework for detecting author spamicity by modeling review deviation," *Expert Systems with Applications* (112), pp. 148–155.

Liu, Y., Pang, B., and Wang, X. 2019. "Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph," *Neurocomputing* (366), pp. 276–283.

Luca, M. 2016. "Reviews, reputation, and revenue: The case of Yelp. com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* (12-016).

Luca, M. and Zervas, G. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science* (62:12), pp. 3412–3427.

Ma, Y. J. and Lee, H.-H. 2014. "Consumer responses toward online review manipulation," *Journal of Research in Interactive Marketing* (8:3), pp. 224–244.

Marciano, J. 2021. *Fake online reviews cost $152 billion a year. Here's how e-commerce sites can stop them*. World Economic Forum. Accessed January 7, 2023.

Mayzlin, D., Dover, Y., and Chevalier, J. 2014. "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review* (104:8), pp. 2421–2455.

McCluskey, M. 2022. *Inside the War on Fake Consumer Reviews*. Time. Accessed January 7, 2023.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. 2013. "What yelp fake review filter might be doing?," in: *Proceedings of the international AAAI conference on web and social media,* pp. 409–418.

Munzel, A. 2016. "Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus," *Journal of Retailing and Consumer Services* (32), pp. 96–108.

Navon, D. and Miller, J. 2002. "Queuing or sharing? A critical evaluation of the single-bottleneck notion," *Cognitive Psychology* (44:3), pp. 193–251.

Nickerson, R. S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of General Psychology* (2:2), pp. 175–220.

OpenAI 2023. *GPT-4 Technical Report*. arXiv: 2303.08774.

Ott, M., Cardie, C., and Hancock, J. 2012. "Estimating the Prevalence of Deception in Online Review Communities," in: *Proceedings of the 21st International Conference on World Wide Web,* WWW '12. Lyon, France: Association for Computing Machinery, pp. 201–210.

Ott, M., Cardie, C., and Hancock, J. T. 2013. "Negative deceptive opinion spam," in: *Proceedings of the 2013 Conference of the NA Chapter of the Association for Computational Linguistics,* pp. 497–501.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. 2011. "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, Oregon, USA, 2011, pp. 309–319.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. 2022. "Training language models to follow instructions with human feedback," in: *Neural Information Processing Systems,* vol. 35, pp. 27730–27744.

Paul, H. and Nikolaev, A. 2021. "Fake Review Detection on Online E-Commerce Platforms: A Systematic Literature Review," *Data Mining and Knowledge Discovery* (35:5), pp. 1830–1881.

Plotkina, D., Munzel, A., and Pallud, J. 2020. "Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews," *Journal of Business Research* (109), pp. 511–523.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. 2019. "Language models are unsupervised multitask learners," *OpenAI blog* (1:8), p. 9.

Rayana, S. and Akoglu, L. 2015. "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* KDD '15. New York, NY, USA: Association for Computing Machinery, pp. 985–994.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-g., and Jansen, B. J. 2022. "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services* (64), p. 102771.

Sandulescu, V. and Ester, M. 2015. "Detecting singleton review spammers using semantic similarity," in: *Proceedings of the 24th international conference on World Wide Web,* Florence, Italy: Association for Computing Machinery, pp. 971–976.

Senter, R. and Smith, E. A. 1967. *Automated readability index*. Tech. rep. Cincinnati Univ OH.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., et al. 2019. *Release strategies and the social impacts of language models*. arXiv: 1908.09203.

Sun, H., Morales, A., and Yan, X. 2013. "Synthetic review spamming and defense," in: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 1088–1096.

Tang, T., Fang, E., and Wang, F. 2014. "Is Neutral Really Neutral? The Effects of Neutral User-Generated Content on Product Sales," *Journal of Marketing* (78:4), pp. 41–58.

Tricomi, P. P., Tarahomi, S., Cattai, C., Martini, F., and Conti, M. 2022. *Are we all in a truman show? spotting instagram crowdturfing through self-training*. arXiv: 2206.12904.

TripAdvisor 2021. *Review Transparency Report*. Tripadvisor Media Center. Accessed January 7, 2023.

Vana, P. and Lambrecht, A. 2021. "The Effect of Individual Online Reviews on Purchase Likelihood," *Marketing Science* (40:4), pp. 708–730.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention is All you Need," in: *Advances in Neural Information Processing Systems,* vol. 30. Curran Associates, Inc.

Wang, X., Sanders, S. P., and Sanders, G. L. 2021. "Examining the Impact of Yelp's Elite Squad on Users' Following Contribution," in: *CIS 2021 Proceedings. 23,* pp. 1–16.

Wickens, C. D., Hollands, J. G., Banbury, S., and Parasuraman, R. 2015. *Engineering Psychology and Human Performance,* Psychology Press.

Wu, Y., Ngai, E. W., Wu, P., and Wu, C. 2020. "Fake online reviews: Literature review, synthesis, and directions for future research," *Decision Support Systems* (132), p. 113280.

Xu, Y., Zhang, Z., Law, R., and Zhang, Z. 2020. "Effects of online reviews and managerial responses from a review manipulation perspective," *Current Issues in Tourism* (23:17), pp. 2207–2222.

Yao, Y., Viswanath, B., Cryan, J., Zheng, H., and Zhao, B. Y. 2017. "Automated Crowdturfing Attacks and Defenses in Online Review Systems," in: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,* CCS '17. Dallas, Texas, USA: Association for Computing Machinery, pp. 1143–1158.

Yoo, K.-H. and Gretzel, U. 2009. "Comparison of Deceptive and Truthful Travel Reviews," in: *Information and Communication Technologies in Tourism 2009,* W. Höpken, U. Gretzel, and R. Law (eds.). Vienna: Springer Vienna, pp. 37–47.

Zhang, D., Zhou, L., Kehoe, J. L., and Kilic, I. Y. 2016. "What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews," *Journal of Management Information Systems* (33:2), pp. 456–481.

Zhang, J. 2019. "What's yours is mine: exploring customer voice on Airbnb using text-mining approaches," *Journal of Consumer Marketing* (36:5), pp. 655–665.

Zhang, M. and Luo, L. 2023. "Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp," *Management Science* (69:1), pp. 25–50.

Zhang, M., Wei, X., and Zeng, D. D. 2020. "A matter of reevaluation: incentivizing users to contribute reviews in online platforms," *Decision Support Systems* (128), p. 113158.

Zhang, T., Li, G., Cheng, T., and Lai, K. K. 2017. "Welfare economics of review information: Implications for the online selling platform owner," *International Journal of Production Economics* (184), pp. 69–79.

Zhang, W., Du, Y., Yoshida, T., and Wang, Q. 2018. "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Information Processing & Management* (54:4), pp. 576–592.

Zhao, Y., Yang, S., Narayan, V., and Zhao, Y. 2013. "Modeling consumer learning from online product reviews," *Marketing Science* (32:1), pp. 153–169.

Zhuang, M., Cui, G., and Peng, L. 2018. "Manufactured opinions: The effect of manipulating online product reviews," *Journal of Business Research* (87), pp. 24–35.