Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

AiGen-FoodReview: A Multimodal Dataset of Machine-Generated Restaurant Reviews and Images on Social Media

Alessandro Gambetti, Qiwei Han
Nova School of Business and Economics
Rua da Holanda 1
2775-405 Carcavelos, Lisbon, Portugal
gambetti.alessandro@novasbe.pt, qiwei.han@novasbe.pt

Abstract

Online reviews in the form of user-generated content (UGC) significantly impact consumer decision-making. However, the pervasive issue of not only human fake content but also machine-generated content challenges UGC's reliability. Recent advances in Large Language Models (LLMs) may pave the way to fabricate indistinguishable fake generated content at a much lower cost. Leveraging OpenAI's GPT-4-Turbo and DALL-E-2 models, we craft **AiGen-FoodReview**, a multimodal dataset of 20,144 restaurant review-image pairs divided into authentic and machine-generated. We explore unimodal and multimodal detection models, achieving 99.80% multimodal accuracy with FLAVA. We use attributes from readability and photographic theories to score reviews and images, respectively, demonstrating their utility as handcrafted features in scalable and interpretable detection models with comparable performance. This paper contributes by open-sourcing the dataset and releasing fake review detectors, recommending its use in unimodal and multimodal fake review detection tasks, and evaluating linguistic and visual features in synthetic versus authentic data.

Introduction

Online reviews on digital platforms are known to significantly influence consumer decisions and trust in products and services. These reviews, integrating real-life experiences, not only offer insights into product quality but also carry economic implications (Duan, Gu, and Whinston 2008; Wu et al. 2015). Complementing textual reviews, hybrid reviews that contain both textual and visual content become increasingly prevalent on social media and enhance the richness of user-generated content (UGC), contributing to the overall impact on consumer perception. In particular, there is an emerging research focus on the synergy of texts and images in online reviews. Studies suggest that hybrid content is often perceived as more informative and helpful than textual content (Wu, Wu, and Wang 2021). For example, research shows that in the context of restaurant reviews, images can have more predictive power for business outcomes than text alone (Zhang and Luo 2023).

While online reviews are vital for consumer decision-making, the credibility of UGC is increasingly compromised by the prevalence of fake reviews. These deceptive reviews, whether overly positive or negative, are intended to manipulate consumer perceptions and distort the online marketplace (Crawford et al. 2015; Paul and Nikolaev 2021). The restaurant industry, in particular, has seen instances where even marginal increases in online ratings can lead to significant revenue growth, further incentivizing the generation of fake reviews (Luca 2016). Platforms such as Amazon, TripAdvisor and Yelp have implemented countermeasures to filter suspicious content. For example, Luca and Zervas (2016) found that 16% of restaurant reviews on Yelp are filtered.

Traditionally, fake reviews have been manually crafted, often by organized fake review farms (He, Hollenbeck, and Proserpio 2022; McCluskey 2022). However, the landscape is evolving with the advancements in Large Language Models (LLMs), which are becoming indistinguishable at producing text from human writing. The accessibility and low cost of these models, such as GPT-based models, pose a new challenge in the proliferation of machine-generated fake reviews (Gambetti and Han 2023). These developments are not limited to text; models like DALL-E are capable of creating realistic images that can accompany fake reviews, adding a new dimension to the challenge of identifying fabricated content (Ramesh et al. 2021). This situation requires a multifaceted strategy to detect and counteract machine-generated fake content. As visual content manipulation becomes more sophisticated, distinguishing genuine from fabricated imagery is crucial in the broader fight against misinformation. Thus, a comprehensive multimodal approach, encompassing both textual and visual elements, is essential for a robust defense against misinformation in digital spaces.

This paper presents **AiGen-FoodReview**, a novel multimodal dataset comprising fake restaurant reviews and images, created using OpenAI models GPT-4-Turbo and Dall-E-2. This dataset, with 20,144 review-image pairs, is a pioneering effort in assembling a comprehensive repository of multimodal machine-generated customer reviews on social media. We analyze this dataset by examining textual attributes like readability, complexity, and perplexity, as well as visual attributes grounded in photographic theory, such as brightness and compositional elements. Our analysis reveals notable differences between machine-generated and authentic content, with machine-generated reviews displaying higher complexity and the images showing distinct

brightness, saturation, and colorfulness.

To address the challenge of separating machine-generated content from authentic material, we developed and optimized several detection models. Our experiments included both unimodal and multimodal machine learning models, focusing on handcrafted features derived from text and images, as well as deep learning models that utilize raw data. In particular, our multimodal FLAVA model, applied to raw data, achieved F1-score of 99.80%. However, it is worth mentioning that models based on handcrafted features also demonstrated strong performance, offering a viable option when scalability and interpretability are key considerations.

This study contributes to the field by introducing a publicly available multimodal dataset of machine-generated fake reviews and images. We also provide a suite of detectors, optimized for identifying such content, and make these tools available to the community. We recommend employing this dataset for the following tasks:

- **Task 1**: Detecting fake restaurant reviews in both unimodal and multimodal contexts.
- **Task 2**: Analyzing and comparing linguistic and visual features of synthetic versus authentic text and image data.

Related Work

Generative AI and Large Language Models

Generative AI systems, capable of producing novel content, have evolved significantly in recent years. Early advancements include Generative Adversarial Networks (GANs), which employ a generator and discriminator in an adversarial training process (Goodfellow et al. 2014), and Variational Autoencoders (VAEs), focusing on probabilistic mappings (Kingma and Welling 2022; Mirza and Osindero 2014). Meanwhile, Pixel Recurrent Neural Networks (PRNN) further extended this concept to sequential image generation (van den Oord, Kalchbrenner, and Kavukcuoglu 2016). These models essentially laid the groundwork for subsequent developments in content generation.

The introduction of transformer architecture marked a turning point in generative AI, particularly influencing the development of LLMs (Vaswani et al. 2017). These models have demonstrated remarkable proficiency in producing human-like text and images, reflecting significant progress in the field (Chen et al. 2021; Kasneci et al. 2023; Koh, Fried, and Salakhutdinov 2023).

LLMs are trained on extensive data to understand and predict patterns, outputting not only indistinguishable human-like text but also generating images via multimodal architectures (Chen et al. 2021; Kasneci et al. 2023; Koh, Fried, and Salakhutdinov 2023). Unimodal LLMs, exemplified by BERT and RoBERTa, focus primarily on encoding textual information for various predictive tasks. Decoder-focused models like the GPT series excel in generating new text, utilizing a masking strategy to enhance predictive accuracy (Devlin et al. 2019; Liu et al. 2019; Brown et al. 2020; Radford et al. 2019). In contrast, multimodal LLMs integrate different modalities, such as text and images, either through fusion-free or fusion-based architectures. Fusion-free models like CLIP and ALIGN align modality-specific representations via contrastive training, while fusion models like FLAVA and ALBEF use explicit cross-attention mechanisms for intermodal interaction (Radford et al. 2021; Jia et al. 2021; Singh et al. 2022; Li et al. 2021).

Moreover, LLMs are increasingly interacted with through prompt engineering, a technique allowing users to guide the generation of responses for specific tasks. This approach has made LLM applications like ChatGPT and Gemini widely accessible for various tasks, including synthetic data generation (White et al. 2023; Zhou et al. 2023; Wu et al. 2023; Veselovsky et al. 2023; Veselovsky, Ribeiro, and West 2023). Lastly, LLMs have shown particular effectiveness in synthetic data generation, which includes text production and, more pertinently, the generation of fake reviews. The democratization of these technologies raises concerns about their potential misuse in creating deceptive online content (Gambetti and Han 2023).

Fake Reviews on Social Media and Online Markets

The proliferation of fake reviews on social media and online markets has become a significant issue. Motivated by monetary gains, retailers and online platforms manipulate reviews to influence consumer behavior, affecting market dynamics and trust (Crawford et al. 2015; Gössling, Hall, and Andersson 2018; Lee, Qiu, and Whinston 2018; Paul and Nikolaev 2021; He, Hollenbeck, and Proserpio 2022). By distorting the informativeness of content, fake reviews have profound implications on consumer decision-making. They foster uncertainty and distrust among consumers, negatively affecting their purchasing intentions (Agnihotri and Bhattacharya 2016; Zhao et al. 2013; Zhang et al. 2017; DeAndrea et al. 2018; Filieri, Alguezaui, and McLeay 2015; Munzel 2016; Xu et al. 2020; Zhuang, Cui, and Peng 2018).

An emerging concern is the use of LLMs for generating fake reviews. The ability of these models to produce authentic-seeming content at scale poses new challenges for online platforms in identifying and mitigating such deceptive practices. This development calls for advanced detection methods and a reevaluation of current strategies to preserve the integrity of online review ecosystems (Gambetti and Han 2023). As machine-generated content becomes more sophisticated, the detection of fake reviews requires not only traditional text analysis but also an understanding of the nuances introduced by AI-generated content. This underscores the need for continuous research and development in AI detection methodologies, emphasizing the importance of both unimodal and multimodal approaches to effectively identify synthetic content.

Dataset Methodology

Raw Data Collection

We leveraged the New York City SafeGraph restaurant mobility data from 2019 to June 2022 as a basis for selecting restaurants to collect reviews (https://www.safegraph.com/). SafeGraph is a company that tracks customer traffic of point-of-interest (POIs) such as restaurants via smartphone signals (Wi-Fi, GPS, and Bluetooth). We selected New York City as a geographical area as it offers a diverse set of international

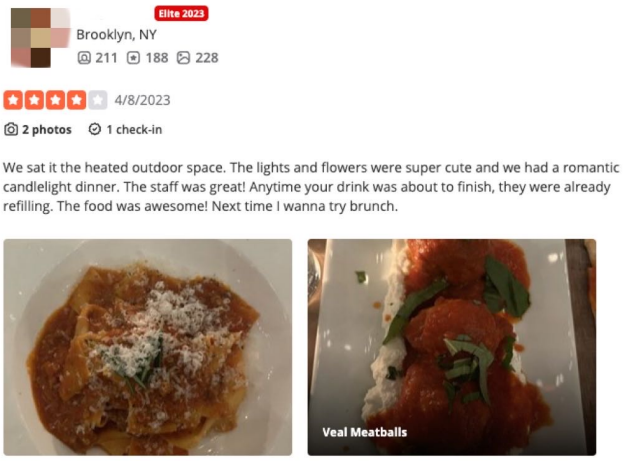

Figure 1: Example of how a scraped review (with images attached by the same user) is displayed on Yelp. The red label indicates the user's elite status. Other variables include user location, user number of friends, number of previous reviews posted, and number of images posted. Username and image were anonymized and blurred.

cuisines in a global setting, ensuring cultural heterogeneity for our research.

From the initial set of 9,200 restaurants, we scraped all English-written reviews from Yelp, a prominent platform for discovering and reviewing local businesses. This effort yielded 447,377 textual reviews from the same period. Alongside each review, we collected associated ratings, user elite status, and downloaded accompanying images, if any. The relationship between reviews and images was carefully mapped, with each review potentially linked to multiple images. Figure 1 shows an example of a scraped review displayed on Yelp.

Data Processing

Our processing involved two key steps. Initially, we focused on reviews with at least one attached image, ensuring a multimodal dataset. For reviews linked to multiple images, we randomly sampled a single image to pair with the text. This random sampling was crucial to maintain diversity and avoid bias in image representation. We then narrowed down to reviews posted by Yelp's elite users, considering these as more reliable sources. This choice is made with two considerations: firstly, elite status on Yelp is a mark of verified and consistent contribution, indicative of trustworthy reviews (Zhang, Wei, and Zeng 2020; Wang, Sanders, and Sanders 2021); secondly, Yelp's filtering system, while effective, may not fully capture all fake reviews (Mukherjee et al. 2013), and selecting only elite reviews adds an additional layer of authenticity. Consequently, we propose that elite reviews offer the closest proxy to genuine customer feedback. Finally, the textual data was cleaned to remove HTML tags and non-ASCII characters, resulting in a dataset of 21,143 elite reviews, each paired with at least one image.

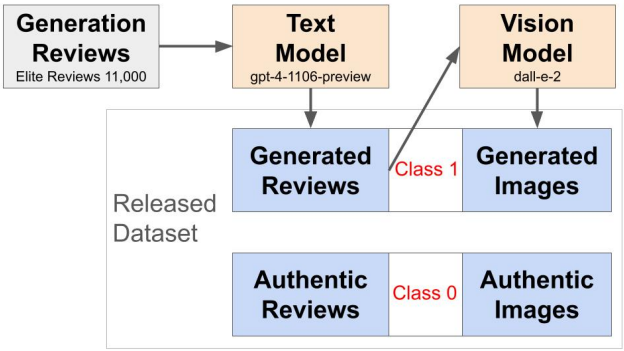

Figure 2: Diagram of the data generation methodology. Elite reviews are used as a source of information in the prompt to query the GPT-4-Turbo model to generate a fake review. Next, the *generated* fake review is used as information to produce a related synthetic image. Finally, review-image pairs, both *authentic* and *generated*, were aggregated into the final dataset to form the negative class (*Class 0*) and positive class (*Class 1*), respectively.

Multimodal Fake Reviews and Images Generation

To generate the multimodal dataset, we employed GPT-4-Turbo for text generation (OpenAI 2023) and Dall-E-2 for image creation (Ramesh et al. 2022). These state-of-the-art models were selected for their advanced capabilities in generating realistic and contextually relevant content. The dataset generation process involved several carefully designed steps, as illustrated in Figure 2. Importantly, we define the following terminology: (1) *generation* are the reviews used as input as the source of information to give contextual knowledge to the language model to generate synthetic text, (2) *generated* refers to the generated output from the language and vision models (both reviews and images), and (3) *authentic* refers to the reviews and images aggregated as ground truth later. Figure 2 shows a visual summary diagram of the data generation methodology. The steps are explained as follows.

Firstly, we randomly divided the 21,143 reviews between 11,000 examples to be used for *generation* and 10,143 as *authentic* to create a binary dataset between the two classes. It is reasonable to separate *generation* data and data to be aggregated later as *authentic* because of self-containment. For example, assuming separation had not been performed, there would be a probability that *generated* reviews would be similar to *generation* reviews. Therefore, in a train-test split scenario, it could be likely that a detector would learn wrong representations.

Secondly, for each review in *generation*, we queried the *gpt-4-1106-preview* model with the following selected prompt, which exactly corresponds to the default prompt embedded into the fake review generator tool that OpenAI provided in its playground as of early 2023 (Gambetti and Han 2023)[1]:

[1]https://platform.openai.com/examples/default-restaurant-review

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

Other hyperparameters such as the *temperature* and the *top_p* were kept as default, except for the *max_length* of the generated output set to 512 tokens to avoid output truncation.

Thirdly, we prompted the *dall-e-2* standard model with the *generated* reviews to generate synthetic images of size 256x256 based on synthetic text. All the hyperparameters were kept as default. Since *dall-e-2* accepts a maximum of 1,000 characters, prompt truncation was performed at that cutoff value. The reason we used *dall-e-2* instead of the most recent *dall-e-3* is two-fold: (1) *dall-e-3* is 2.5 times more costly than *dall-e-2* with prices of $0.040/image (standard model, size 1024×1024) and $0.016/image (standard model, size 256x256), respectively, and (2) we noticed from our experiments that *dall-e-3* outputs less realistic and more cartoonish images than *dall-e-2*. We validated that by prompting *dall-e-3* with the same hyperparameters as for *dall-e-2* but adding *style* equal to *natural* to avoid generating hyper-real images. In Figure 3 we show 3 samples of generated images from the two models using the same prompt out of a total of 30 pairs generated. Validation was done manually by the authors.

Finally, because of content moderation imposed by OpenAI, we recorded 999 errors in generating images, meaning that 999 *generated* reviews could not be "translated" into images (Error code: 400, "content_policy_violation"). Thus, we aggregated to the 10,001 *generated* review-image pairs (*Class 1*) the 10,143 *authentic* reviews and related images (*Class 0*). In total, the final multimodal review dataset counts 20,144 review-image pairs.

Generation Cost Breakdown. We here report the cost incurred for the steps illustrated above. As of January 2024, *gpt-4-1106-preview* prices input text at $0.01/1K tokens and output text at $0.03/1K tokens. Here, we counted the average number of input and output tokens using the popular Python package "tiktoken". Whereas, *dall-e-2* (standard 256x256) prices a single image generation at $0.016. Arithmetically, we can break the *TOTAL* cost as:

$$\begin{aligned} TC &= (0.01 * IC + 0.03 * OC) * M \\ VC &= 0.16 * M \\ TOTAL &= TC + VC \end{aligned} \tag{1}$$

TC stands for the text cost using a textual model, *VC* stands for the image cost, $IC = 188$ stands for the average number of input tokens (average length in *generation*), and $OC = 157$ stands for the average number of output tokens (average length in *generated*). Finally, $M = 11,000$ is the number of *generation* reviews. In total, TC amounts to about $72, while VC amounts to about $176, for a *TOTAL* cost of about $248.

It is important to highlight that machine-generated reviews are relatively cheap to generate, as one machine-generated review costs, on average, about $0.00654 ($TC/M$). For the sake of comparison, one investigation from Which, a UK consumer-right group, reported the cost of buying human-generated fake reviews at $18 per review (Dean 2021).

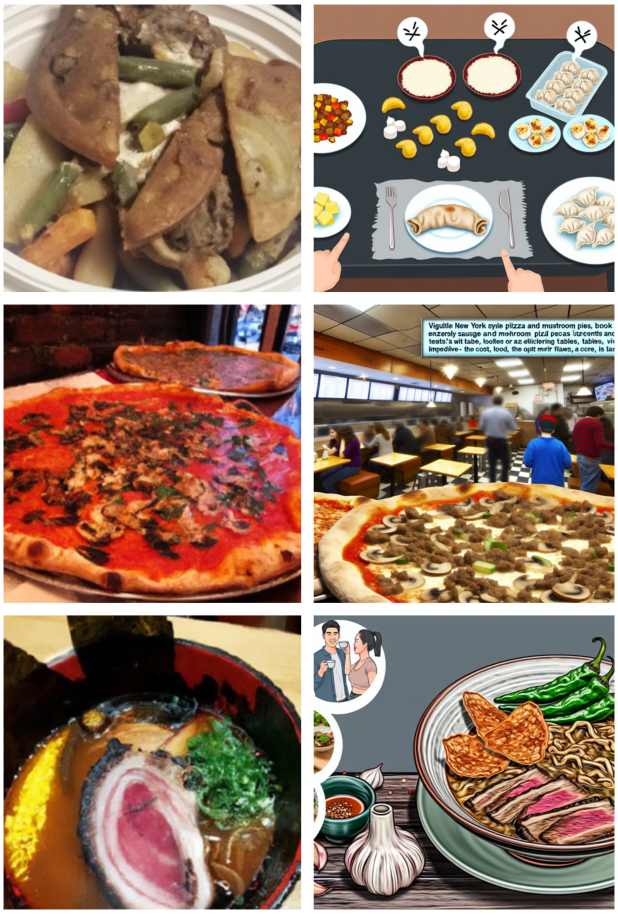

Figure 3: Generated Images from Dall-E-2 (left column) and Dall-E-3 (right column) for the same prompt.

Dataset Summary Statistics

We provide a concise dataset summary statistics at review, image, and restaurant levels for the restaurants represented in *authentic* and in *generated*. As for terminology clarification, we interchangeably refer to the *authentic* reviews as *Class 0*, and to the *generated* reviews as *Class 1*, respectively.

Review Level. We report that average ratings of the subset reviews, *i.e.*, those in *generation* and *authentic*, (4.07±0.90) are slightly higher in comparison to those of the raw dataset (3.96±1.36), as measured on a 1-5 Likert scale. But, no rating differences across *Class 0* and *Class 1* were recorded, with average ratings of 4.074±0.91 and 4.073±0.90, respectively.

Next, following Gambetti and Han (2023), we calculated review text statistics by mining readability, complexity, and perplexity handcrafted attributes. Here, we scored each review with the (1) Automated Readability Index (ARI), which measures an approximate representation of the US

grade level needed to comprehend the text (Senter and Smith 1967); (2) the Fleisch Readability index (FR), in which higher scores indicate text that is easier to read (Flesch 1948); (3) the number of difficult words (DW) present in the Dale-Chall word list, which approximately contains 3,000 words of difficult understanding (Dale and Chall 1948); (4) the Gunning Fog Index (GFI), which has a similar meaning as for ARI; (5) the reading time (RT), which calculates the reading time assuming 14.69ms per character; (6) the average words per sentence (WPS); and (7) perplexity (PPL). Here, for each review, we calculated PPL as the exponential weighted average of the negative log-likelihoods of a word sequence W as $PPL(W) = \exp\left[-\frac{1}{t}\sum_1^t \log p(w_i|w_{<i})\right]$. To do so, we implemented a zero-shot 125M parameters GPTNeo model, which is an open-source replication of the GPT-3 model (Black et al. 2022). We utilized ANOVA to assess the statistical differences among variables. In Table 1, we describe the variables, and in Table 2 we report the text summary statistics.

Results indicate that: *generated* reviews are more complex to read as compared to *authentic* ones, as measured by higher ARI and GFI, lower FR, and a larger number of DW. Also, *generated* reviews score an average lower perplexity, which is coherent with LLMs optimization logic, as such models are generally optimized by minimizing perplexity in the self-supervised learning paradigm. Finally, *authentic* reviews have longer sequences (higher WPS), which translates into longer reading time (RT).

Image Level. We scored each image with the attributes illustrated by Gambetti and Han (2022) to calculate statistics of photographic attributes per class, including their developed food aesthetic score for food images only (FA). To do so, we sampled 20,000 labeled images from the Yelp official dataset[2], split the data into 80% train, and 20% test, and fine-tuned a ViT-B/16 model to classify food versus non-food images, achieving 97.63% accuracy on the test set. Then, we applied FA to food images, ranging from 0 (low aesthetics) to 1 (high aesthetics). Next, we considered the following photographic attributes: (1) color attributes such as brightness (BRI), saturation (SAT), contrast (CON), clarity (CLA), warmth (WAR), and colorfulness (COL); (2) figure-ground relationship attributes such as size difference (SD), color difference (CD), and texture difference (TD); and (3) image composition attributes such as diagonal dominance (DD), rule of thirds (ROT), horizontal/vertical physical visual balance (HPVB, VPVB), and horizontal/vertical color visual balance (HCVB, VCVB). We utilized ANOVA to assess the statistical differences among variables. In Table 1 we describe the variables, and in Table 2 we show the summary statistics.

Results indicate that: *generated* images tend to score a higher aesthetics score (FA). As for color attributes, *generated* images are brighter (higher BRI), more saturated (higher SAT), clearer (higher CLA), warmer (higher WAR), and more colorful (higher COL). Next, as for figure-ground relationship, both *authentic* and *generated* images show comparable size differences (SD) and texture differences (TD) between the figure and the ground. However, *generated* images tend to be more separated in terms of color difference between foreground and background (CD). Finally, although statistically significant, as for image composition, no clear patterns could be evinced across the two classes, showing comparable diagonal dominance (DD), rule of thirds (ROT), and physical visual balance (HPVB, VPVB).

[2]https://www.yelp.com/dataset

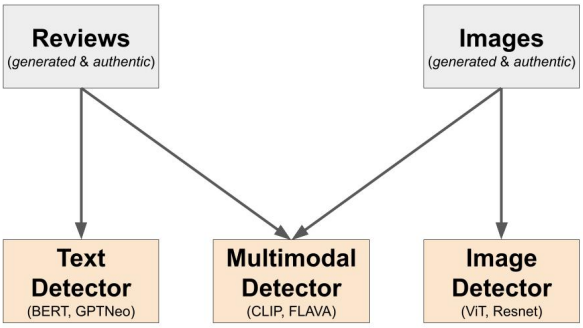

Figure 4: Representation of unimodal and multimodal feature combinations for detection models.

Restaurant Level. There are 3,238 restaurants represented in the dataset, with an average of about 6 reviews per restaurant. We then queried the official Yelp Fusion API to gather information about the regional cuisines represented in the sample (https://fusion.yelp.com/). American, Italian, and Japanese are the most popular cuisines with market shares of 18.84%, 13.22%, and 5.78%, respectively. Next, in terms of price levels, which indicate the average meal price per person, 1,941 restaurants (59.94%) are priced in the range \$11–\$30, as denoted by \$\$. In descending order, 452 restaurants are priced between \$31–\$60 (\$\$\$), 372 restaurants are priced below \$10 (\$), and 119 over \$61 (\$\$\$\$). Finally, the average of the restaurants' average rating amounts to 3.80 ± 0.51.

Experiments

We trained unimodal and multimodal binary fake review detectors to test how they perform on our crafted dataset. Firstly, we randomly split the generated dataset into 60% train, 20% validation, and 20% test. Secondly, we trained unimodal models on text and image data separately, and finally trained multimodal models on text and image data together (see Figure 4). We also report the performance of open-source models as benchmarks. Evaluation metrics for comparison include the accuracy score, precision score, recall score, and F1-score.

Text Detectors

We benchmarked with the official OpenAI RoBERTa model for fake text detection (Solaiman et al. 2019). We applied it directly to the test set. Next, we fine-tuned a 110M parameters BERT and a 125M parameters GPTNeo (Black et al. 2022) for binary classification. We trained with AdamW

Metric	Description	Range
Text		
ARI	US grade level needed for comprehension. Higher = Harder to read.	$[0, \infty)$
FR	Higher = Easier to read.	$[0, 100]$
DW	Count of hard words not in top 3,000 common words. Higher = Harder to read.	$[0, \infty)$
PPL	Words predictability. Higher = Lower predictability.	$[0, \infty)$
GFI	US grade level needed. Higher = Harder to read.	$[0, \infty)$
RT	Reading time in ms. Higher = Longer.	$(0, \infty)$
WPS	Average number of words per sentence. Higher = Purportedly longer.	$(0, \infty)$
Image		
FA	Food aesthetics score (Gambetti and Han 2022). Higher = Higher aesthetics.	$[0, 1]$
BRI	Average of V of the HSV image representation. Higher = Brighter.	$[0, 180]$
SAT	Average of S of the HSV image representation. Higher = More saturared.	$[0, 180]$
CON	Standard deviation of V of the HSV image representation. Higher = More contrast.	$(0, \infty)$
CLA	% of normalized V pixels that exceed 0.7 of HSV. Higher = Clearer.	$[0, 1]$
WAR	% of H $<$ 60 or $>$ than 220 of HSV. Higher = Warmer.	$[0, 1]$
COL	Departure from a grey-scale image. Higher = More colorful.	$[0, \infty)$
SD	Normalized number of pixel difference between figure and ground. Higher = More distinct.	$[-1, 1]$
CD	RGB Color Euclidean distance difference between figure and ground. Higher = More distinct.	$[0, \infty)$
TD	Absolute difference between the foreground and background edge density. Higher = More distinct.	$[0, \infty)$
DD	Manhattan distance between salient region and each diagonal. Higher = More diagonally dominant.	$(-\infty, 0]$
ROT	Minimum distance from salient region center to each 4 intersection points. Higher = More centered.	$(-\infty, 0]$
HPVB	Horizontal physical visual balance via mirrored split image. Higher = More balanced.	$(-\infty, 0]$
VPVB	Vertical physical visual balance via mirrored split image. Higher = More balanced.	$(-\infty, 0]$
HCVB	Horizontal color balance: mirrored Euclidean cross-pixels. Higher = More balanced.	$(-\infty, 0]$
VCVB	Vertical color balance: mirrored Euclidean cross-pixels. Higher = More balanced.	$(-\infty, 0]$

Table 1: Description of the text and image handcrafted variables mined in the paper. Except for PPL, Text features were calculated using the TEXTSTAT Python package. Documentation with formulae available at: https://pypi.org/project/textstat/, as well as in our formula supplement in our GitHub repository at: https://github.com/iamalegambetti/aigen-foodreview/. Whereas, Image feature calculations were adapted from Gambetti and Han (2022), Section 4.2.

with a learning rate of 1e-4, a batch size of 16, truncation and padding at 512 tokens, and early stopping at 5 epochs. We trained on the training set and evaluated accuracy after each epoch on the validation set. Finally, we reported the results of the test set.

Image Detectors

By directly applying it to the test set, we benchmarked with an open-source CvT-13 model trained to detect AI-generated images by Horbatko (2023), who optimized on ArtiFact (Rahman et al. 2023), a generalist dataset for synthetic image detection. Then, we fine-tuned a ViT-B/16 and a ResNet-50. We trained with equal methodology to text models: AdamW with a learning rate of 1e-4, batch size of 16, and early stopping at 5 epochs. We followed the default image-processing steps for both models in their original papers.

Multimodal Detectors

To the best of our knowledge, no robust benchmarks exist for text-image multimodal deepfake detection. Thus, we could not benchmark prior models with our dataset. Hence, we fine-tuned pre-trained CLIP (based on a ViT-B/16), a fusion-free multimodal model (Radford et al. 2021); and FLAVA, a late-fusion multimodal model (Singh et al. 2022). Both were pre-trained by learning representations from image and text pairs through contrastive training. As for CLIP, we extracted the separate image and text-learned representations and concatenated them. On top, a linear MLP was added as a classification head. As for FLAVA, we added a linear MLP on the jointly learned representation at the last layer as a classification head. Without freezing the backbone models, we trained with the same hyperparameters and methodology as for unimodal models: AdamW with a learning rate 1e-4, a batch size of 16, truncation at 512 tokens (except for CLIP, which truncates inputs at 76 tokens), and early stopping at 5 epochs.

Handcrafted Features Detectors

We evaluated whether handcrafted textual and image features contribute to detecting machine-generated reviews. Using the features mined in the *Dataset Summary Statistics* section, we optimized Logistic Regression models (LR) and Random Forest (RF) on: (1) text features only, *i.e.*, readability, complexity, and perplexity features, (2) image features only, *i.e.*, photographic attributes, and (3) a multimodal aggregation including both sets of features. We trained on the training set, and reported results on the test set. Hyperparameters were kept as default. Results are reported in Table 3. Finally, given the explainability nature of RF and the handcrafted features, we applied SHAP on the test set to check the feature importance of the most impactful features

Metric	Authentic	Generated	F-statistic
Text			
ARI	6.84 (3.00)	12.20 (1.90)	22,883***
FR	79.93 (9.26)	57.78 (8.15)	32,424***
DW	19.77 (13.97)	39.06 (15.23)	8,783***
PPL	55.56 (41.43)	38.19 (11.31)	1,638***
GFI	7.71 (2.37)	12.29 (1.66)	25,309***
RT	9.93 (6.82)	11.96 (4.44)	626***
WPS	14.10 (5.39)	19.05 (3.15)	6,303***
Image			
FA	0.07 (0.15)	0.35 (0.34)	4,604***
BRI	132.79 (30.41)	144.51 (33.73)	671***
SAT	112.65 (35.79)	117.35 (42.45)	1,822***
CON	59.16 (11.66)	69.81 (12.48)	3,919***
CLA	0.30 (0.18)	0.41 (0.18)	1,822***
WAR	0.28 (0.21)	0.45 (0.21)	3,183***
COL	150.11 (18.67)	160.88 (14.49)	2,089***
SD	0.31 (0.29)	0.29 (0.32)	18.74***
CD	79.78 (52.21)	93.94 (57.78)	332***
TD	4.06 (11.11)	3.87 (4.64)	2.69
DD	-32.76 (33.77)	-36.92 (33.52)	76.78***
ROT	-75.65 (23.29)	-79.67 (15.16)	209.67***
HPVB	-9.09 (8.11)	-9.00 (7.46)	0.78
VPVB	-6.41 (5.77)	-6.36 (5.34)	0.40
HCVB	-0.54 (0.14)	-0.48 (0.02)	1,977***
VCVB	-0.54 (0.15)	-0.48 (0.03)	1,598***

Table 2: Summary statistic of text and image attributes. Average values are reported with standard deviation in brackets. *p <.05, **p<.01, ***p<.001.

in making predictions. For general reference, SHAP (SHapley Additive exPlanations) is a widely used interpretability framework in machine learning, offering a unified and theoretically grounded approach to quantify the contribution of each feature in a model's output (Lundberg and Lee 2017).

Results & Discussion

We summarize the binary classification report on the test set in Table 3. Results indicate that restaurant *generated* content is separable from *authentic* content both in unimodal and multimodal experiments. Benchmarks' performance is not sufficient to achieve satisfactory accuracy. Here, the official OpenAI RoBERTa detector performed worse than random guessing (<50%) in the unimodal text setting, and Horbatko (2023) vision model achieves only 73.67% accuracy in the unimodal image setting. In contrast, all the fine-tuned models achieved about 99% performance in all the metrics evaluated. With an accuracy comparable to BERT (99.38%), GPTNeo achieves the best accuracy and F1-score for text models with values of 99.68% and 99.68%, respectively. As for unimodal image models, ResNet-50 has comparable accuracy to ViT-B/16 (99.21% and 99.16%, respectively). As for multimodal models, FLAVA outperformed CLIP by +1.34% with an accuracy of 99.80%, and all the other unimodal models. Finally, as for models trained on handcrafted features, RF trained on multimodal features achieves the best performance (98.96% accuracy). This is followed by

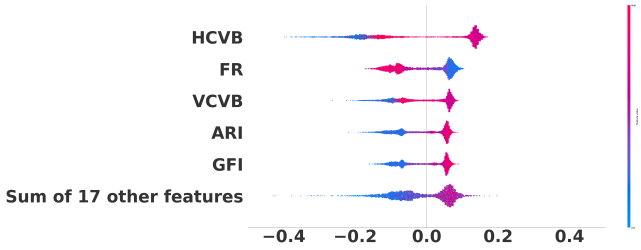

Figure 5: SHAP evaluation: top 5 influential features to predict *generated* reviews in descending order. The concentration of red dots in the $x > 0$ quadrant, with the blue ones on the $x < 0$ quadrant, implies a positive correlation with the target variable, and vice versa. For example, ARI is positively correlated, while FR is negatively correlated.

the RF on image features (98.19% accuracy) and LR on multimodal features (96.30% accuracy). We then applied SHAP explainability to the multimodal RF model. Figure 5 shows the top 5 influential features in predicting a *generated* review-image pair. From these results, both horizontal and vertical color visual balance (HCVB, VCVB) have a higher weight, as well as Flesch Reading index (FR), Automated Readability Index (ARI), and Gunning Fog Index (GFI).

Overall, all the fine-tuned models, except for LR trained on handcrafted image features, achieved robust performance in the hold-out dataset. However, multimodal FLAVA achieves the overall best performance both in accuracy and F1-score, suggesting that detection from multimodal input can be an effective alternative when possible. However, CLIP performed worse, conjecturing that CLIP's forced truncation at 76 textual input tokens can be a limiting factor during optimization. For reference, input for the other text models was truncated at 512 tokens. Inherently, CLIP's convergence after 20 epochs might be a consequence stemming from this limitation. A valid alternative to deep learning models was the adoption of standard machine learning models trained on handcrafted features. The latter achieved comparable performance, but with faster training and inference time, enabling quicker experimentation. These factors should be taken into account in real-world scenarios in which scalability plays a central role in project development. Also, training machine learning models on handcrafted features enhances their interpretability and explainability, facilitating a clearer understanding of the factors influencing their predictions. In this regard, we showed that SHAP is a valuable framework for providing explainability to a machine-generated fake reviews detector based on handcrafted features.

Limitations

This work is not without limitations. Firstly, our training sample is restricted to only include reviews for restaurants from the New York City area, which may not be representative of all reviews from different regions. Secondly, we only adopted one generative model: GPT-4-Turbo and Dall-E-2 for text and vision tasks, respectively, because these OpenAI-based models received wide acceptance and pop-

Model	Type	Conv@Epoch	Accuracy %	Precision %	Recall %	F1-score %
LR (ours)	Handcrafted/Text	-	94.99	95.14	94.96	95.05
RF (ours)	Handcrafted/Text	-	95.46	95.81	95.21	95.51
LR (ours)	Handcrafted/Image	-	78.24	78.62	78.42	78.52
RF (ours)	Handcrafted/Image	-	98.19	98.24	98.19	98.21
LR (ours)	Handcrafted/Multi	-	96.30	96.65	96.03	96.34
RF (ours)	Handcrafted/Multi	-	98.96	99.31	98.63	98.97
Solaiman et al. (2019)	Text	-	44.94	22.57	03.52	06.09
BERT (ours)	Text	1	99.38	99.51	99.27	99.39
GPTNeo (ours)	Text	6	99.68	99.51	**99.85**	99.68
Horbatko (2023)	Image	-	73.67	88.01	68.80	77.23
ViT-B/16 (ours)	Image	3	99.16	99.46	98.88	99.17
ResNet-50 (ours)	Image	8	99.21	**99.90**	98.55	99.22
CLIP (ours)	Multimodal	20	98.46	98.34	98.63	98.48
FLAVA (ours)	Multimodal	1	**99.80**	**99.90**	99.71	**99.80**

Table 3: Classification report on the test set. Numbers in bold indicate best performance. Conv@Epoch indicates the epoch in which weights were saved after early stopping.

ularity (Ghassemi et al. 2023). However, alternative generative models could have been used, such that the dataset may also contain content generated beyond OpenAI-based models. Thirdly, we only selected one prompt template for both text and image generation. However, it is likely that generated content may be affected by the diversified prompt structure, resulting in varied complexity. Fourthly, we previously highlighted how the ratings of subset reviews in *generation* and *authentic* are slightly more positive than those in the raw dataset, such that detection performance may change across rating classes (1-5 stars). However, we briefly tested how the best FLAVA detector performed across test set reviews of different ratings, noticing that F1-scores are substantially equal, ranging from the lowest of 99.58% (4 stars) to the highest of 100% (3 stars). Lastly, we assumed that elite reviews would be authentic content, because Yelp intentionally promotes their elite squad program to allow members in the program to share more reliable experiences (see https://www.yelp.com/elite for details on *"Real people. Real reviews.®"*). Still, we cannot exclude the possibility that members of the elite squad could be secretly paid to circulate fake reviews. However, it is beyond the scope of this study to identify such fraudulent practices.

Conclusion

In this paper, we introduced **AiGen-FoodReview**, a multimodal dataset of machine-generated reviews and images. The dataset consists of 20,144 review-image pairs, divided into 10,143 *authentic* and 10,001 *generated* pairs. *Authentic* reviews were scraped from Yelp, while *generated* reviews and images were generated using GPT-4-Turbo and Dalle-E-2, respectively, at a cost of approximately $248, significantly lower than those incurred by hiring human workers (see Dean (2021)), underscoring the need for future research on whether machine-generated content may become more prevalent than the human-generated one. Our analysis indicates that GPT-4-Turbo *generated* reviews are linguistically more complex, and Dalle-E-2 *generated* images exhibit enhanced brightness, saturation, and clarity compared to authentic ones. We also optimized fake review-image detectors, achieving reasonable performance on the test set. Given the mass adoption of LLMs, we conjecture a surge in machine-generated content, potentially impacting user experiences and trust in social media platforms. In conclusion, our dataset serves as an open-source benchmark for studying multimodal machine-generated reviews and images.

Ethical Impact and FAIR

This dataset release aligns with ethical standards to ensure integrity in research. Privacy and confidentiality have been rigorously maintained through anonymization. No private information about Yelp users has been collected. We release the dataset with an MIT license. Researchers are advised to engage in the responsible utilization of this dataset, as well as not using the content of this dataset to spread misinformation on the web. We adhere to the **FAIR** principles. The dataset is **Findable** on Zenodo through a unique DOI identifier: 10.5281/zenodo.10511456. The dataset files are **Accessible**, as they can be retrieved without incurring any charges. Ensuring **Interoperability**, the dataset is provided in simple and standardized formats (CSV for reviews and JPG for images). Finally, the dataset is designed to be highly **Re-usable**. The inclusion of a metadata.txt file serves to document information about the variables, offering comprehensive insights into the dataset's structure. In addition to the Zenodo repository at: https://zenodo.org/records/10511456, we release our work on GitHub at: https://github.com/iamalegambetti/aigen-foodreview, including both trained unimodal and multimodal detectors.

Acknowledgments

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).

References

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

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; Pieler, M.; Prashanth, U. S.; Purohit, S.; Reynolds, L.; Tow, J.; Wang, B.; and Weinbach, S. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. arXiv:2204.06745.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. arXiv:2107.03374.

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): 1–24.

Dale, E.; and Chall, J. S. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*, 37–54.

Dean, G. 2021. Websites are selling fake reviews 'in bulk' to Amazon merchants, a report found. One site offered 1,000 reviews for $11,000. https://www.businessinsider.com/fake-amazon-reviews-for-sale-buy-merchants-amazons-choice-2021-2. Accessed: 2024-01-03.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

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

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: 174–185.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3): 221–233.

Gambetti, A.; and Han, Q. 2022. Camera eats first: exploring food aesthetics portrayed on social media using deep learning. *International Journal of Contemporary Hospitality Management*, 34(9): 3300–3331.

Gambetti, A.; and Han, Q. 2023. Dissecting AI-Generated Fake Reviews: Detection and Analysis of GPT-Based Restaurant Reviews on Social Media. In *Proceedings of the International Conference on Information Systems*, 8. Aisnet.

Ghassemi, M.; Birhane, A.; Bilal, M.; Kankaria, S.; Malone, C.; Mollick, E.; and Tustumi, F. 2023. ChatGPT one year on: who is using it, how and why? *Nature*, 624(7990): 39–41.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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): 484–503.

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

Horbatko, L. 2023. AI Image Detector. https://github.com/guyfloki/ai-image-detector. Accessed: 2024-01-07.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Koh, J. Y.; Fried, D.; and Salakhutdinov, R. 2023. Generating Images with Multimodal Language Models. arXiv:2305.17216.

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): 393–416.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 9694–9705. Curran Associates, Inc.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

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): 3412–3427.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc.

McCluskey, M. 2022. Inside the War on Fake Consumer Reviews. https://time.com/6192933/fake-reviews-regulation. Accessed: 2024-01-07.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784.

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*, volume 7, 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: 96–108.

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

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): 1830–1881.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

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): 9.

Rahman, M. A.; Paul, B.; Sarker, N. H.; Hakim, Z. I. A.; and Fattah, S. A. 2023. ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection. arXiv:2302.11970.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Senter, R.; and Smith, E. A. 1967. Automated readability index. Technical report, Amrl-Tr. Aerospace Medical Research Laboratories.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. FLAVA: A Foundational Language And Vision Alignment Model. arXiv:2112.04482.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; McCain, M.; Newhouse, A.; Blazakis, J.; McGuffie, K.; and Wang, J. 2019. Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203.

van den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel Recurrent Neural Networks. arXiv:1601.06759.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Veselovsky, V.; Ribeiro, M. H.; Arora, A.; Josifoski, M.; Anderson, A.; and West, R. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. arXiv:2305.15041.

Veselovsky, V.; Ribeiro, M. H.; and West, R. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv:2306.07899.

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*, 1–16.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382.

Wu, C.; Che, H.; Chan, T. Y.; and Lu, X. 2015. The economic value of online reviews. *Marketing Science*, 34(5): 739–754.

Wu, R.; Wu, H.-H.; and Wang, C. L. 2021. Why is a picture 'worth a thousand words'? Pictures as information in perceived helpfulness of online reviews. *International Journal of Consumer Studies*, 45(3): 364–378.

Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; and Tang, Y. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5): 1122–1136.

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): 2207–2222.

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): 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: 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: 69–79.

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

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2023. Large Language Models Are Human-Level Prompt Engineers. arXiv:2211.01910.

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

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. In the Limitations section, it was described that focusing on New York City may be a population-specific dataset.
   (e) Did you describe the limitations of your work? Yes. A dedicated subsection has been included.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. In the ethical and FAIR section.
   (g) Did you discuss any potential misuse of your work? Yes. In the ethical and FAIR section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. In the ethical and FAIR section.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA.
   (b) Have you provided justifications for all theoretical results? NA.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA.
   (e) Did you address potential biases or limitations in your theoretical framework? NA.
   (f) Have you related your theoretical results to the existing literature in social science? NA.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. Data and code have been made open-source.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. In the Text Detectors, Image Detectors and Multimodal Detectors Sections.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? No, because we used basic evaluation metrics such as accuracy, precision, recall and F1-score, which are implicitly suitable for the evaluation of a machine learning model.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? Yes. MIT License. Section Ethical Impact and FAIR.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes. Section Ethical Impact and FAIR.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. This section was included.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No, because of the simple structure of the data, which was released both in Zenodo and GitHub.
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? NA.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.
   (d) Did you discuss how data is stored, shared, and deidentified? NA.